DOCUMENT RESUME

ED 321 318 CS 507 209

AUTHOR	Studdert-Kennedy, Michael, Ed.
TITLE	Status Report on Speech Research, July-December 1989.
INSTITUTION	Haskins Labs., New Haven, Conn.
SPONS AGENCY	National Inst. of Child Health and Human Development (NIH), Bethesda, Md.; National Inst. of Health (DHHS), Bethesda, MD. Biomedical Research Support Grant Program.; National Inst. of Neurological and Communicative Disorders and Stroke (NIH), Bethesda, Md.; National Science Foundation, Washington, D.C.
REPORT NO	SR-99/100
PUB DATE	89
CONTRACT	NO1-HD-5-2910
GRANT	BNS-8520709; HD-01994; NS07237; NS13617; NS13870; NS18010; NS24655; NS25455; RR-05596
NOTE	221p.; For the January-June 1989 status report, see CS 507 208.
PUB TYPE	Collected Works - General (020) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Auditory Perception; Communication Research; *Language Processing; Language Research; Orthographic Symbols; *Phonology; Reading; *Speech; *Speech Communication
IDENTIFIERS	Speech Perception; Speech Research; Vocalization

ABSTRACT

One of a series of semiannual reports, this publication contains 12 articles which report the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. The titles of the articles and their authors are as follows: "Coarticulatory Organization for Lip-rounding in Turkish and English" (Suzanne E. Boyce); "Long Range Coarticulatory Effects for Tongue Dorsum Contact in VCVCV Sequences" (Daniel Recasens); "A Dynamical Approach to Gestural Patterning in Speech Production" (Elliot L. Saltzman and Kevin G. Munhall); "Articulatory Gestures as Phonological Units" (Catherine P. Browman and Louis Goldstein); "The Perception of Phonetic Gestures" (Carol A. Fowler and Lawrence D. Rosenblum); "Competence and Performance in Child Language" (Stephen Crain and Janet Dean Fodor); "Cues to the Perception of Taiwanese Tones" (Hwei-Bing Lin and Bruno H. Repp); "Physical Interaction and Association by Contiguity in Memory for the Words and Melodies of Songs" (Robert G. Crowder and others); "Orthography and Phonology: The Psychological Reality of Orthographic Depth" (Ram Frost); "Phonology and Reading: Evidence from Profoundly Deaf Readers" (Vicki L. Hanson); "Syntactic Competence and Reading Ability in Children" (Shlomo Bentin and others); and "Effect of Emotional Valence in Infant Expressions upon Perceptual Asymmetries in Adult Viewers" (Catherine T. Best and Heidi Freya Queen). An appendix lists DTIC and ERIC numbers for publications in this series since 1970. (SR)

ED321318

Haskins Laboratories Status Report on

Speech Research

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. DADOURIAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CS507209

SR-99/100
JULY-DECEMBER 1989

BEST COPY AVAILABLE

Haskins
Laboratories
Status Report on

Speech Research

SR-99/100
JULY-DECEMBER 1989

NEW HAVEN, CONNECTICUT

Distribution Statement

Editor
Michael Studdert-Kennedy

Production Staff
Yvonne Manning
Zefang Wang

This publication reports the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications.

Distribution of this document is unlimited. The information in this document is available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.

Correspondence concerning this report should be addressed to the Editor at the address below:

Haskins Laboratories
270 Crown Street
New Haven, Connecticut 06511-6695

This Report was reproduced on recycled paper

Acknowledgment

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development

Grant HD-01994
Contract NO1-HD-5-2910

National Institutes of Health

Biomedical Research Support Grant RR-05596

National Science Foundation

Grant BNS-8520709

National Institute of Neurological and Communicative Disorders and Stroke

Grant NS 13870
Grant NS 13617
Grant NS 18010
Grant NS 24655
Grant NS 07237
Grant NS 25455

Investigators

Arthur Abramson*
Peter J. Alfonso*
Thomas Baer*
Eric Bateson*
Fredericka Bell-Berti*
Catherine T. Best*
Susan Brady*
Catherine P. Browman
Franklin S. Cooper*
Stephen Crain*
Robert Crowder*
Lois G. Dreyer*
Alice Faber†
Laurie B. Feldman*
Janet Fodor*
Anne Fowler
Carol A. Fowler*
Louis Goldstein*
Carol Gracco†
Vincent Gracco
Vicki L. Hanson*
Katherine S. Harris*
Leonard Katz*
Rena Arens Krakow*
Andrea G. Levitt*
Alvin M. Liberman*
Isabelle Y. Liberman*
Diane Lillo-Martin*
Leigh Lisker*
Anders Löfqvist*
Virginia H. Mann*
Ignatius G. Mattingly*
Nancy S. McGarr*
Richard S. McGowan
Patrick W. Nye
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Elliot Saltzman
Donald Shankweiler*
Michael Studdert-Kennedy*
Koichi Tsunoda[1]
Michael T. Turvey*
Douglas Whalen

Technical/Administrative Staff

Philip Chagnon
Alice Dadourian
Michael D'Angelo
Betty J. DeLise
Vincent Gulisano
Donald Hailey
Raymond C. Huey*
Marion MacEachron*
Yvonne Manning
Joan Martinez
Lisa Mazur
William P. Scully
Richard S. Sharkany
Zefang Wang
Edward R. Wiley

Students*

Jennifer Aydelott
Dragana Barac
Paola Bellabarba
Melanie Campbell
Sandra Chiang
André Cooper
Margaret Hall Dunn
Terri Erwin
Elizabeth Goodell
Joseph Kalinowski
Deborah Kuglitsch
Simon Levy
Katrina Lukatela
Diana Matson
Gerald W. McRoberts
Pamela Mermelstein
Salvatore Miranda
Maria Mody
Weijia Ni
Mira Peter
Nian-qi Ren
Christine Romano
Arlyne Russo
Richard C. Schmidt
Jeffrey Shaw
Caroline Smith
Jennifer Snyder
Robin Seider Story
Rosalind Thornton
Mark Tiede
Qi Wang
Yi Xu

*Part-time
†NRSA Training Fellow
[1]Visiting from University of Tokyo, Japan

Contents

Coarticulatory Organization for Lip-rounding in Turkish and English
Suzanne E. Boyce 1

Long Range Coarticulatory Effects for Tongue Dorsum Contact in VCVCV Sequences
Daniel Recasens 19

A Dynamical Approach to Gestural Patterning in Speech Production
Elliot L. Saltzman and Kevin G. Munhall 38

Articulatory Gestures as Phonological Units
Catherine P. Browman and Louis Goldstein 69

The Perception of Phonetic Gestures
Carol A. Fowler and Lawrence D. Rosenblum 102

Competence and Performance in Child Language
Stephen Crain and Janet Dean Fodor 118

Cues to the Perception of Taiwanese Tones
Hwei-Bing Lin and Bruno H. Repp 137

Physical Interaction and Association by Contiguity in Memory for the Words and Melodies of Songs
Robert G. Crowder, Mary Louise Serafine, and Bruno H. Repp 148

Orthography and Phonology: The Psychological Reality of Orthographic Depth
Ram Frost 162

Phonology and Reading: Evidence from Profoundly Deaf Readers
Vicki L. Hanson 172

Syntactic Competence and Reading Ability in Children
Shlomo Bentin, Avital Deutsch, and Isabelle Y. Liberman 180

Effect of Emotional Valence in Infant Expressions upon Perceptual Asymmetries in Adult Viewers
Catherine T. Best and Heidi Fréya Queen 195

Appendix 211

Status Report on

Speech Research

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 1-18

Coarticulatory Organization for Lip-rounding in Turkish and English*

Suzanne E. Boyce†

INTRODUCTION

Theories of coarticulation in speech have taken as an axiom the notion that, by coarticulating segments, a speaker is aiding the efficiency of his or her production (Liberman & Studdert-Kennedy, 1977). Although discussion of the forces affecting coarticulation has tended to concentrate on articulatory and/or perceptual pressures operating within particular sequences of segments (Beckman & Shoji, 1984; Martin & Bunnell, 1982; Ohala, 1981; Recasens, 1985), there is a growing body of cross-linguistic work exploring the influence of language-particular phonological structure on coarticulation (Keating, 1988; Lubker & Gay, 1982; Magen, 1984; Manuel, 1990; Öhman, 1966; Perkell 1986, among others). These studies have generally been concerned with the interaction of coarticulation and segment inventory, or coarticulation and the properties of some particular segment; the question of how coarticulation interacts with phonological rules has been relatively neglected (but cf. Cohn, 1988). Phonological rules, for instance, determine the typical structure of words in a language; we might speculate that languages with different constraints on the possible sequencing of segments pose different challenges to the articulatory planner, and thus that speakers of these languages would vary in the way they implement coarticulation. To take an example, speakers of Turkish, a vowel harmony language with strict rules for the possible sequencing of rounded and unrounded vowels, might feel more pressure to employ rounding coarticulation than English speakers, whose language freely combines rounded and unrounded vowels.

Rounding coarticulation for sequences of rounded and unrounded vowels in English has been extensively studied. A number of studies have shown that for strings of two rounded vowels separated by non-labial consonants, e.g., /utu/ or /ustu/, both EMG and lip protrusion movement traces show double peaks coincident with the two rounded vowels plus an intervening dip or trough in the signal (Engstrand, 1981; Gay, 1978; MacAllister, 1978; Perkell, 1986). (A schematized version of this pattern, representing EMG from the orbicularis oris muscle for the utterance /utu/, is illustrated in Figure 1.) This result has been the focus of a good deal of controversy in recent years, primarily because different theories of coarticulation tend to treat it in different ways. For instance, much previous work on the control mechanisms underlying anticipatory coarticulation has centered on the predictions of one class of models, the "look-ahead" or "feature-spreading" models (Benguerel & Cowan, 1974; Daniloff & Moll, 1968; Henke, 1966). Generally, these models view coarticulation as the migration of features from surrounding phones. In the most explicit form of this type of model, Henke's (1966) computer implementation of articulatory synthesis, the articulatory planning component scans upcoming segments and implements features as soon as preceding articulatorily compatible segments make it possible to do so.[1] In the case of /utu/ or /ustu/, the non-labial consonants separating the vowels are made with the tongue and presumably do not conflict with simultaneous lip-rounding. Thus, the fact that troughs occur is a problem for the look-ahead model, because the model would normally predict that the rounding feature for the second vowel

The research in this paper was supported by NIH grants NS-13617 and BRS RR-05596 to Haskins laboratories. Suggestions, comments and criticism supplied by Katherine Harris, Louis Goldstein, Michael Studdert-Kennedy, Ignatius Mattingly, Fredericka Bell-Berti, Sharon Manuel, Rena Krakow, Marie Huffman, Joe Perkell and John Westbury, and the *JASA* review process are gratefully acknowledged. Advice on Turkish linguistics was provided by Jaklin Kornfelt and Engin Sezer.

would spread onto the preceding consonant, producing a continuous plateau of rounding from vowel to vowel.

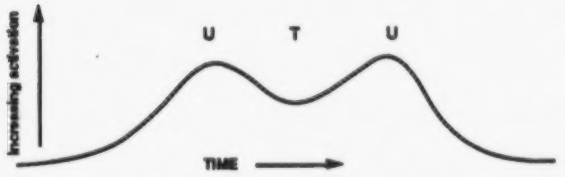


Figure 1. Schematized version of "trough" pattern, representing EMG from the orbicularis muscle for the utterance /utu/.

Explanations of the trough results have varied widely. Following a more general suggestion of Kozhevnikov and Chistovich (1965), Gay (1978) proposed that the trough represents the resetting of a "syllable-sized articulatory unit," and that coarticulation is allowed to take place only within that unit. Some evidence against this explanation was provided by Harris and Bell-Berti (1984), who found no sign of a trough in sequences such as /uhu/ and /u?u/. Addressing himself to the sequences typically used in these experiments, Engstrand (1981) took issue with the assumption that alveolar consonants are compatible with full lip-rounding. He argued instead that rounding as found in the vowel /u/ may interfere with optimal acoustic/aerodynamic conditions for these consonants, and that the trough may result from lip movement towards a less-rounded configuration. That such acoustic/aerodynamic constraints may not hold for all subjects was shown by Gelfer, Bell-Berti and Harris (1989), who reported data from a subject with lip protrusion and EMG Orbicularis Oris Inferior (OOI) activity for /t/. Perkell (1986) hypothesized that a diphthongal pattern of movement for the /u/ vowels (i.e., from a less to a more extreme lip position), might, in addition to acoustic and other constraints on the consonants, reduce the extent of rounding in the vicinity of the intervocalic consonant(s). In his own work, however, he found little evidence for diphthongal behavior in those subjects who showed troughs.

Each of these proposals, it should be noted, can be seen as a modification to the look-ahead class of models, in which features of one segment spread to another segment if context conditions allow it. Alternatively, a class of models known as "coproduction," "frame" or "time-locking" models (Bell-Berti & Harris, 1981; Fowler 1980) assumes that coarticulation results from temporal overlap between independent articulatory gestures belonging to neighboring segments. In these models, lip movement for the the utterance /utu/ involves two overlapping rounding gestures (assuming /t/ is not independently rounded). Thus, the presence of a trough is controlled by the degree of overlap between gesture peaks. If the peaks overlap one another, no trough will be discernible, but if the peaks are temporally separated from one another, the model predicts the occurrence of a trough. Another provision of these models is that gestures associated with a particular segment should show a stable profile across different segmental contexts. (It is acknowledged that characteristic gesture profiles may be affected by stress and possibly other prosodic contexts (Tuller, Kelso, & Harris, 1982). Thus, the temporal extent of coarticulation is predicted by the temporal extent of the gesture. Attempts to test the latter provision of this model by measuring the lag times between the acoustic onset of rounded vowels and related articulatory activity have had varied and sometimes conflicting results, with studies by Bell-Berti and Harris (1979, 1982) and Engstrand (1981) supporting the coproduction prediction of stable lag times, and studies by Lubker (1981), Lubker and Gay (1982), Sussman and Westbury (1981) and Perkell (1986) indicating more variable behavior supportive of the look-ahead view. Thus, although it is unclear how the look-ahead class of models can account for the trough pattern, both types of models remain viable options for explaining coarticulation.

Regardless of which interpretation of the trough pattern is correct, however, the pattern itself has been found in each of the languages so far surveyed, appearing in English (Bell-Berti & Harris, 1974; Gay, 1978; Perkell, 1986), Swedish (Engstrand, 1981; McAllister, 1978), Spanish and French (Perkell, 1986). It is notable that these languages, while differing in such variables as syllable structure, the tendency to diphthongize vowels, and the presence of a phonological contrast between rounded and unrounded vowels, are alike in their tolerance for mixed sequences of rounded and unrounded vowels. It seemed plausible, at least, that the finding of troughs in lip-rounding activity for these languages might be related to this tolerance, and that a language like Turkish, in which words with mixed rounded and unrounded vowels are the exception, might show lip-rounding patterns other than the trough pattern. In particular, it seemed that Turkish

might provide particularly favorable conditions for anticipatory coarticulation of the kind predicted by the look-ahead class of models. In brief, the hypothesis was that Turkish speakers would exhibit plateau patterns of activity for lip-rounding. The experiment described below was part of a larger study designed to test this hypothesis (Boyce, 1988). A second aim of the experiment was to test the degree to which the coproduction model's explicit prediction of stable, independent gestures could be used to predict both English and Turkish movement patterns.

EXPERIMENT

Four speakers of American English and four speakers of Standard Turkish produced similarly structured nonsense words designed to show the presence or absence of troughs in lip-rounding. Corpus words for this purpose consisted of the series /kuktluk/, /kuktuk/, /kukuk/, /kutuk/, /kuluk/. Because arguments concerning the trough pattern often hinge on questions concerning the production of the intervocalic consonants in an unrounded environment (Benguerel & Cowan 1974; Gelfer, Bell-Berti, & Harris 1989), the words /kiktlik/, /kiktik/, /kikik/, /kitik/, /kilik/, were included as controls. Additionally, words with rounded vowels followed by unrounded vowels /kuktlik/, /kuktik/, /kukik/, /kutik/, and /kulik/, and words with unrounded vowels followed by rounded vowels /kiktluk/, /kiktuk/, /kikuk/, /kituk/, and /kiluk/ were included to provide data on single protrusion movements. In the remainder of the paper, words with vowel sequences u-u, ɪ-ɪ, etc. will be referred to as u-u, ɪ-ɪ, u-ɪ, and ɪ-u words. The words with intervocalic KTL, which had the longest vowel-to-vowel intervals, were included to provide the clearest test case for the presence of a trough pattern. Words with shorter intervocalic consonant intervals were included to provide control information on the lip activity patterns for different consonants. The carrier phrase for Turkish speakers was "Bir daha _____ deyiniz" (pronounced as phonetically spelled and meaning 'Say _____ once again'). The English carrier phrase was "Its a _____ again."

English-speaking subjects included one male (AE) and three females (MB, AF and NM), each of whom spoke a variety of General American with no marked regional or dialectal accent. The Turkish speakers included one female (IB) and three males (AT, EG and CK). All spoke similar varieties of Standard Turkish.

Additional facts about Turkish which impinge on the arguments made in this paper have been summarized in the Appendix. For the present, it is sufficient to note that rounding in Turkish operates according to a vowel harmony rule which, in essence, causes sequences of high vowels to acquire the rounding specification of the preceding leftmost vowel. (With minor exceptions, consonants do not participate in this process.) The effect is to produce long strings of rounded or unrounded vowels whose rounding is predictable given the first vowel in the sequence. While vowel harmony is a productive rule for the vast bulk of the lexicon there are numerous exceptions, mainly from Arabic and Persian borrowings. Real word counterparts exist for each of the vowel sequences in the experimental corpus, although u-ɪ and ɪ-ɪ words conform to vowel harmony while ɪ-u and u-ɪ words do not.

For Turkish subject EG, utterances were randomized and the randomized list repeated 15 times. Utterances in later subject runs were blocked, so that utterances were repeated in groups of five tokens (three for MB), utterances with the same vowel combinations were grouped together, and the same order of consonant combinations was repeated for each vowel combination. The order of vowel and consonant combinations was different for each subject, except that one Turkish speaker (IB) and one English speaker (AF) had the same order of presentation.

Although Turkish has final stress, the degree of difference between stressed and unstressed syllables is much less than in English (Boyce, 1978). Therefore, English speakers were encouraged to use equal stress on both syllables of the disyllabic nonsense words, and if equal stress felt unnatural, to place stress on the final rather than the initial syllable. Turkish subjects were given no instructions about stress. All subjects were instructed to speak at a comfortable rate, in a conversational manner.

Instrumentation

Movement data from the nose, upper and lower lip, and jaw were obtained by means of an opto-electrical tracking system, similar to the commonly used Selcom Selspot system. The system consists of infrared light emitting diodes (LED's) attached to the structure of interest. LED position is sensed by a photo-diode within a camera positioned to capture the range of LED movements in its focal plane. The output of this diode is translated by associated electronics into pairs of X and Y coordinate potentials for each LED, each with a maximum frequency response of

500 Hz. Calibration is achieved by moving a diode through a known distance in the focal plane.

LED's were attached to the subject's nose, upper lip, lower lip and jaw with double-sided tape. The nose LED was placed on the bridge of the nose, slightly to the left side, at a point determined to show the least speech-related wrinkling, waggling, etc. LED's were placed just below the vermilion border of the upper lip and just above the vermilion border of the lower lip, in a plane with the nose LED, at a point judged to show the axis of anterior-posterior movement for each articulator. The movement of the subject's skin between the lower lip and chin was observed during production of rounded vowels, and the jaw LED positioned to best reflect anterior-posterior movements of the mandible rather than skin and muscle. Generally this was at the point of the chin or under it, in a plane with the higher LED's.

The LED-tracking camera was positioned at 90 degrees to the left of the subject's sagittal midline, at a camera-to-subject distance (21 inches) that provided a 10-by-10 inch field of view. When centered approximately on the upper/lower lip junction, during maintenance of a position appropriate for bilabial closure, this field is large enough to capture the full range of anterior-posterior LED movement, as well as allowing for some degree of head movement.

A video camera was positioned 90 degrees to subject midline on the subject's right, and focused as narrowly as possible, while continuing to keep all 4 LED's within the field of view. Five subjects were videotaped throughout the experiment: English subjects AF and NM, and Turkish subjects AT and IB. An additional videotape of English subject MB producing the words /kitklik/, kuktluk/, /kiktluk/ and /kuktlik/ was obtained in a separate session.

A simultaneous audio recording of the subject's speech during the experiment was made on a Sennheiser "shotgun" microphone.

The EMG recordings were made with adhesive surface silver-silver chloride electrodes. These were placed just below and above the vermilion border of upper and lower lips, laterally to the midline. According to Blair and Smith (1986), an electrode at this location is likely to pick up relatively more activity from orbicularis oris, and less of nearby muscles, than at other locations along the lip edge. Pick-up from the desired muscles, Orbicularis Oris Inferior (OOI) and Orbicularis Oris Superior (OOS), was checked by having the subject produce repeated /u/ or /i/ vowels several times in succession; if a strong signal was evidenced for /u/ and little or no signal for /i/, the EMG electrode was assumed to be well-placed.

The EMG and movement signals, together with audio and clock signals, were recorded onto a 14-channel FM tape recorder (EMI series 7000). The EMG signals were rectified, integrated over a 5 ms window, and sampled at 200 Hz. Movement signals were also sampled at 200 Hz. The audio channel was filtered at 5000 Hz and sampled at 10000 Hz. By means of the simultaneous clock signals, data from all channels were synchronized to within 2.5 ms.

The signal from the nose LED was numerically subtracted from respective lip and jaw signals to control for changes in baseline due to head movement. Differences in baseline between early and late portions of the experiment remained for some speakers, presumably due to vertical rotational movement of the head, in which the lips and nose, or jaw and nose, moved by different amounts in space. The speakers most affected were English speaker AE, for whom the total horizontal lower lip baseline change was approximately 3 mm, and Turkish speakers IB, EG and CK, for whom the total changes were approximately 3.5, 6.5, and 6 mm respectively. In each case, baseline change reflected movement in the posterior direction. Baseline change for other speakers was within 1 mm of movement. Rotational movement of this type, in which the chin sank gradually toward the base of the neck, was confirmed in the videotape for subject IB (other videotaped subjects showed little baseline change). These baseline changes did not appear to affect the data in any significant way.[2]

Because of recording or calibration problems, the upper lip movement signal and both EMG signals for Turkish subject AT, the EMG OOS signal for Turkish subject EG, and the EMG OOI signal for Turkish subject CK were eliminated from the study. Except for AT, therefore, the full complement of movement signals, and at least one EMG signal, was available for each subject. Recording or calibration problems also caused some of the 15 repetitions (tokens) planned for words in the experimental corpus to be discarded. The upper lip signal level for English subject AE deteriorated after the first block of utterances. Thus, only the first five tokens for this signal are reported.

Two acoustic reference points, or lineups, were identified for each token. The first, the V_1 offset, was defined as the point where the formant structure disappeared from the waveform at the

onset of closure for /k/ or /t/ or the point of sudden amplitude change marking the change between the vowel and the voiced approximant /l/. The second, the V_2 onset, was defined as the release of the consonant occlusion for /k/ and /t/, or the point of amplitude change for /l/. Consonant interval duration measurements consisted of the time between these two points. The audio waveform, movement and EMG signals for each repeti tion of an utterance in the experimental corpus were extracted into a separate computer file. Each file contained a 2000 ms slice of speech with constant dimensions before and after the V_1 offset point.

The main body of movement data reported here comes from the anterior-posterior upper and lower lip signals. These signals are referred to in the text as Upper Lip X (ULX) and Lower Lip X (LLX). Both signals reflect lip protrusion, which is generally acknowledged to be the most reliable single index of lip rounding. However, because rounding may also involve vertical motion of the lips, to narrow the lip aperture, and because vertical movement and protrusion of the lower lip may be affected by movements of the jaw, anterior-posterior jaw (JX) and inferior-superior jaw (JY) and lip signals (ULY, LLY) were examined as well.

As a rule, token-to-token variability was minimal in both movement and EMG signals. Accordingly, much of the presentation in this paper is based on movement and EMG traces produced by ensemble averaging. (Those cases where token-to-token variability was greater than implied by the averaged signal are men tioned in the text.) Signals were ensemble averaged using the acoustic V_1 offset as a lineup point.

Results

Turkish Speakers

Movement and EMG signals were examined separately for Turkish and English subjects, with a view to determining characteristic movement and muscle activity patterns for u-u words. Figures 2 through 5 show the averaged ULX, LLX and EMG traces for /kuktluk/ and /kiktlik/ as produced by the four Turkish speakers AT, IB, EG and CK.

Overall, the /kuktluk/ movement traces for these subjects tended to resemble a plateau, with the protrusion traces being flat or slightly falling over the course of the word. Exceptions to this pattern are the occurrence of a peak, or bump during the consonant interval in the LLX traces for subjects EG and CK, and the slight trough located at the beginning of V_2 in the ULX signal for subject EG. Out of the three Turkish speakers with EMG data (OOS and OOI for IB, OOI for EG and OOS for CK), there was no conspicuous diminution of EMG activity during the consonant interval. The general pattern was unimodal. For IB and CK, there was an early peak on V_1 followed by a long, sustained offset and some indication of increased activity during V_2. For subject EG, the EMG peak was located close to V_2 in /kuktluk/ and to V_1 in all other u-u words. (Movement patterns, however, were similarly plateau-like over EG's different u-u words.)[3]

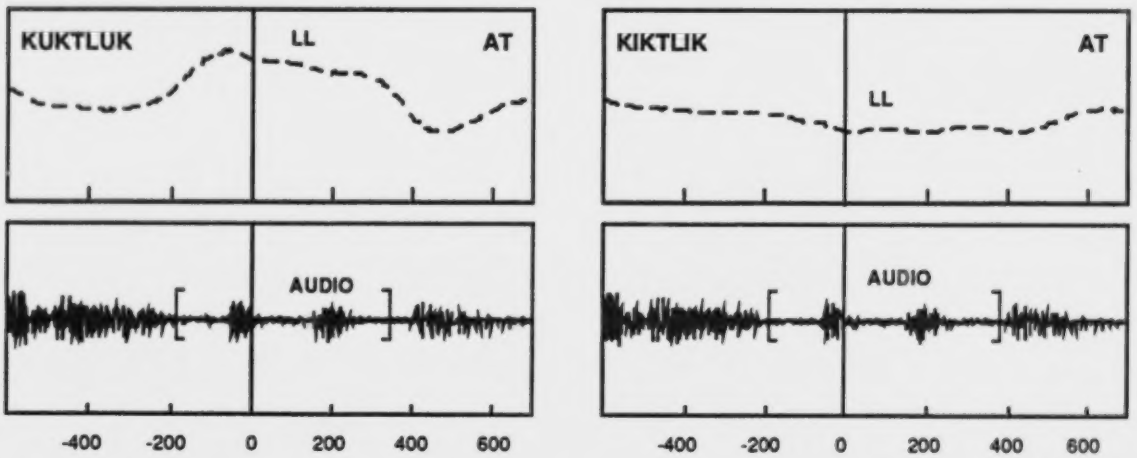


Figure 2. **Averaged lower lip movement traces (dashed line) plus single token acoustic waveform traces, for /kuktluk/ (15 tokens) and /kiktlik/ (15 tokens) as produced by Turkish speaker AT. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for the lower lip (LL) trace is 0-20 mm. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

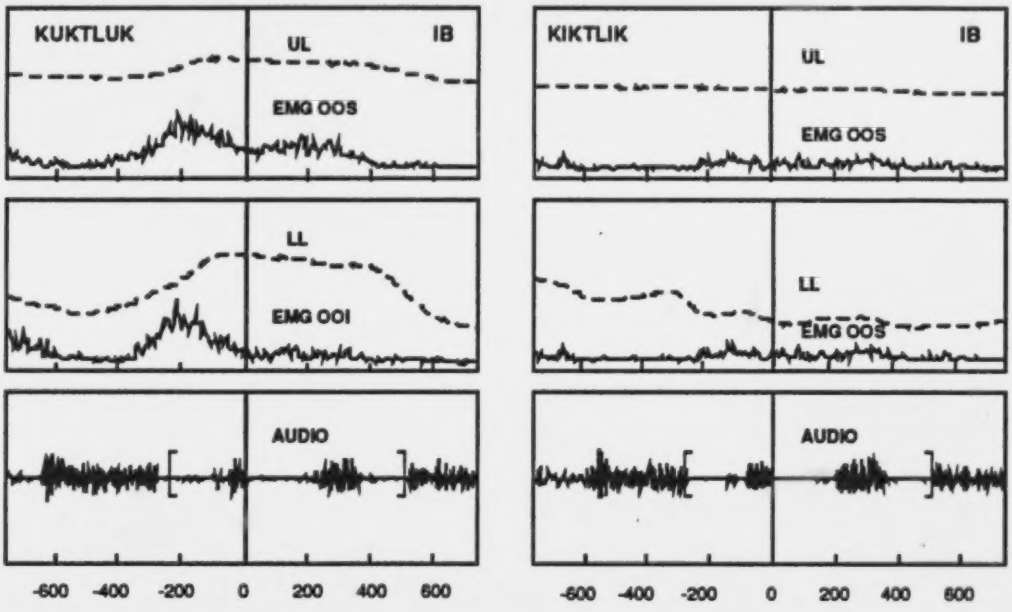


Figure 3. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ (15 tokens) and /kiktlik/ (15 tokens) as produced by Turkish speaker IB. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-20 mm. The vertical scale for both EMG OOS and EMG OOI traces is 200 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

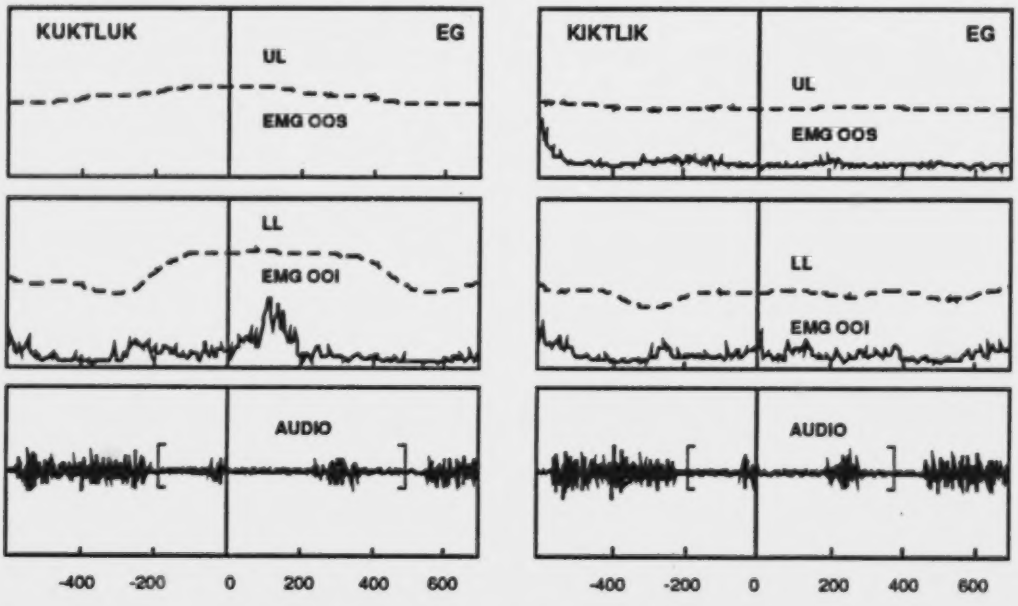


Figure 4. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ (14 tokens) and /kiktlik/ (15 tokens) as produced by Turkish speaker EG. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-20 mm. The vertical scale for EMG OOI traces is 300 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

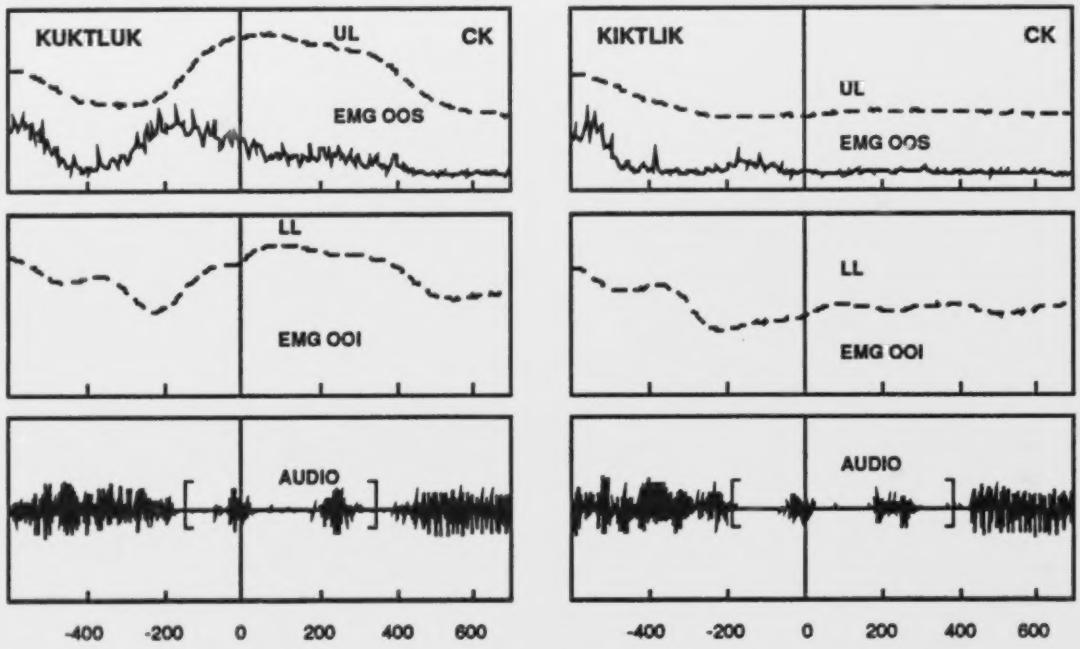


Figure 5. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ (8 tokens) and /kiktlik/ (15 tokens) as produced by Turkish speaker CK. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-10 mm. The vertical scale for EMG OOS traces is 150 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

As noted in the introduction, it is well known that some English speakers may protrude and/or narrow their lips for non-labial consonants, most notably /t/ (Gelfer, Bell-Berti, & Harris, 1989) and /l/ (Brown, 1981; Leidner, 1973). Looking at the /kiktlik/ traces, it appears that Turkish subjects AT, IB and EG do not produce signifi cant independent protrusion during the intervocalic consonants. Although small fluctuations in LLX signals may indicate some degree of active lower lip protrusion, these may also be due to lip relaxation from a retracted position during the flanking /i/ vowels. The strongest degree of movement during the /kiktlik/ consonant interval is seen for subject CK. It is hard to tell if this reflects active movement rather than passive relaxation, however, as CK retracts lip and jaw heavily for the sustained /a/ of "daha" (pronounced [daa]) in the carrier phrase (Boyce, 1988). EMG traces for all subjects during ɪ-ɪ words were flat.

English subjects in this study could be divided into two groups based on the appearance of their horizontal lower lip signals for ʊ-ʊ and ɪ-ɪ words. (Upper lip signals were less clearly differentiated.) Examples of these patterns can be seen in Figures 6 and 7, which show the averaged ULX, LLX and EMG traces for /kuktluk/ and /kiktlik/ as produced by English subjects AE and AF. For both speakers, /kuktluk/ movement traces showed double-peaked trough patterns. A similar double-peaked pattern can be seen for subject AF's EMG OOI and OOS. For subject AE, the EMG OOS trace is clearly double-peaked. The EMG OOI trace also shows two peaks, but with an additional peak between.[4]

The right-hand panels of Figures 5 and 6 show averaged ULX, RLLX and EMG traces for the word /kiktlik/. As can be seen, for subject AF there is little or no movement for either lip during the intervocalic consonant interval, and little or no activity in the EMG signal. Some fluctuation in the movement signal for LLX is present for AE. Comparison with the presumably neutral position of the lips during the schwa vowel from again at the end of the carrier phrase (between 400 and 600 ms after the V_1 offset point) suggests that this may be due to lip retraction during the flanking vowels, with relaxation of the lips during the consonant interval. Alternatively, some small active forward movement of the lower lip and/or jaw may be involved.

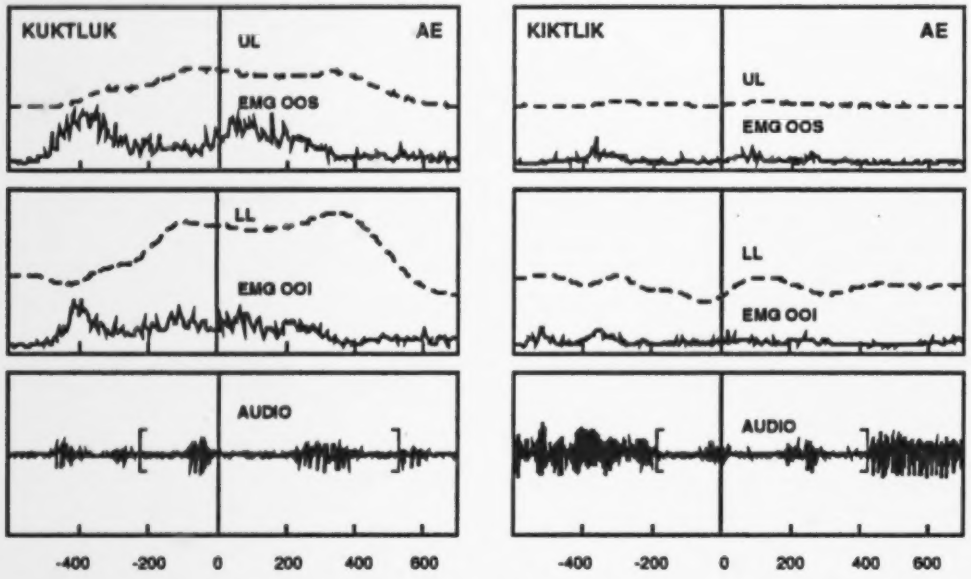


Figure 6. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ and /kiktlik/ as produced by English speaker AE. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. Upwards deflection represents anterior movement. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-20 mm. The vertical scale for both EMG OOS and EMG OOI traces is 200 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The upper lip traces for /kuktluk/ and /kiktlik/ are averaged from 5 tokens. The lower lip trace for /kuktluk/ is averaged from 15 tokens, that for /kiktlik/ from 13 tokens. The horizontal scale is time in ms.**

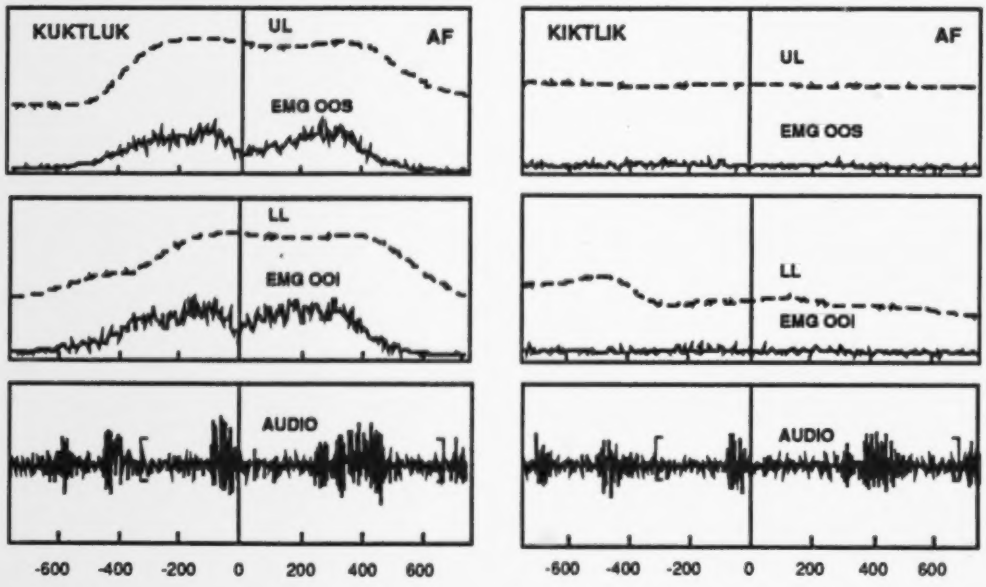


Figure 7. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ (15 tokens) and /kiktlik/ (15 tokens) as produced by English speaker AF. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-20 mm. The vertical scale for both EMG OOS and EMG OOI traces is 500 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

The left and right panels of Figures 8 and 9 show averaged ULX, LLX and EMG traces for English subjects MB and NM. Looking at the /kuktluk/ words, we see that for both MB and NM, the lower lip trace shows three peaks of movement. The first peak is located during the "It's" of the carrier phrase (at approximately 350 ms before the V_1 offset point) and probably indicates protrusion associated with /s/. The central, and largest, peak is located during the intervocalic consonant interval (between the two vertical lines). The third peak is located during the second vowel. At the same time, the EMG patterns for MB and NM's EMG OOI traces show trough patterns like those of English subjects AE and AF (NM's EMG OOS trace, like AE's OOI trace, shows an additional peak after V_1 offset). The upper lip pattern for subject NM also resembles those for subjects AF and AE. Subject MB's upper lip movement pattern contains two peaks, which correspond roughly in time to the central and final peaks of the lower lip trace.

There is less apparent consistency between upper lip, lower lip, and EMG traces for these subjects than for subjects AE and AF. Looking at their /kiktlik/ traces, however, we see that both subjects MB and NM show protrusion in the lower lip signal during the consonant interval. There is also some protrusion in MB's upper lip /kiktlik/ trace. The latter is likely to reflect active forward movement since there is no sign of upper lip retraction on the flanking /i/ vowels.

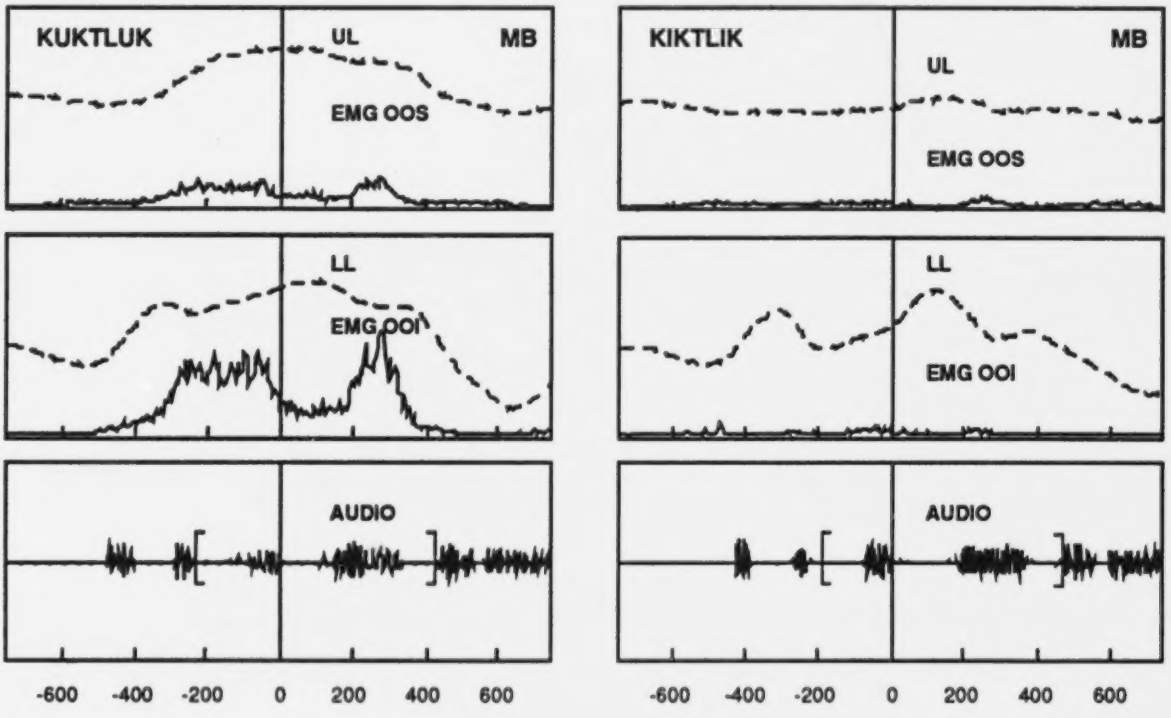


Figure 8. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ (15 tokens) and /kiktlik/ (15 tokens) as produced by English speaker MB. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-15 mm. The vertical scale for both EMG OOS and EMG OOI traces is 800 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

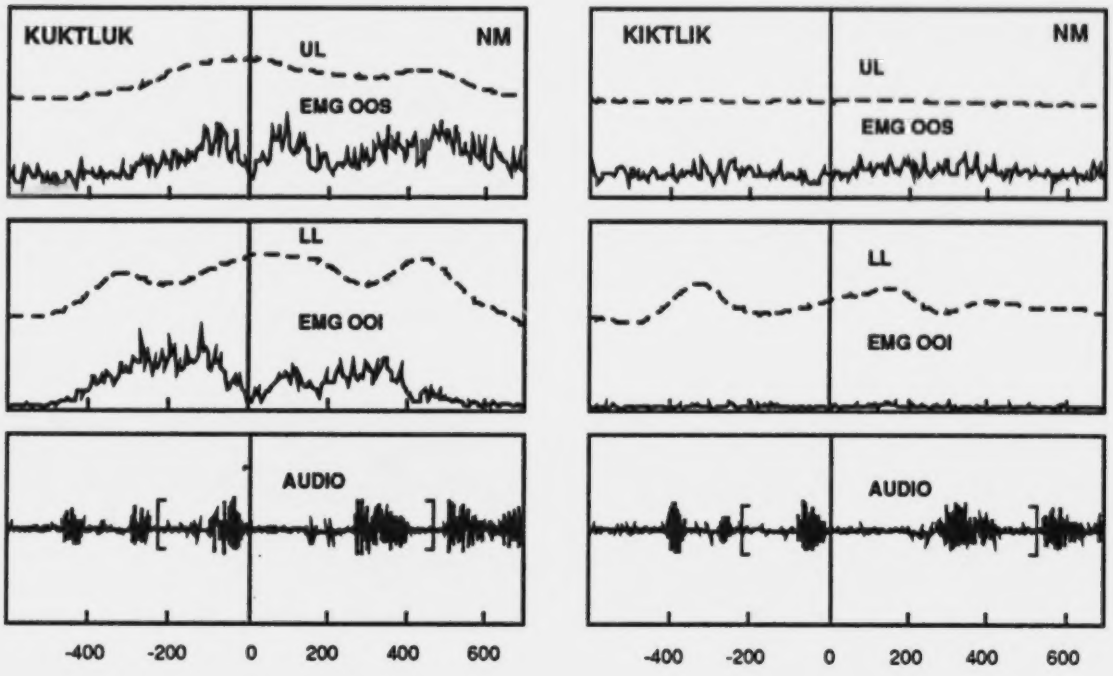


Figure 9. **Averaged movement (dashed line) and EMG (solid line) traces, plus single token acoustic waveform traces, for /kuktluk/ (15 tokens) and /kiktlik/ (14 tokens) as produced by English speaker NM. Upwards deflection represents anterior movement. The vertical line indicates the lineup point for ensemble averaging, which was at the acoustic offset of V_1. The vertical scale for both upper lip (UL) and lower lip (LL) traces is 0-10 mm. The vertical scale for both EMG OOS and EMG OOI traces is 500 μv. Square brackets in the lower panel indicate approximate acoustic boundaries for these words. The horizontal scale is time in ms.**

In Figures 10 and 11, we see the lower lip signal for /kiktlik/ overlaid with that for /kuktluk/ for English subject MB and NM. (Baseline differences between averaged traces have been adjusted when necessary, so as to visually align carrier phrase portions of each trace.) From this, it is clear that the timing of the consonant interval protrusion peak in /kiktlik/ is very similar to the timing of the central peak in these subjects' /kuktluk/ traces. This is most striking for MB, whose protrusion peak in /kiktlik/ was also similar in amplitude to that of /kuktluk/. For NM, the central peak in /kuktluk/ was slightly bimodal, and the protrusion peak in /kiktlik/ is more nearly matched in timing to the second inflection. For both subjects, a similar congruence of peaks can be seen when lower lip /kiktluk/ (shown in Figures 20 and 21) traces are overlaid with /kiktlik/ traces.

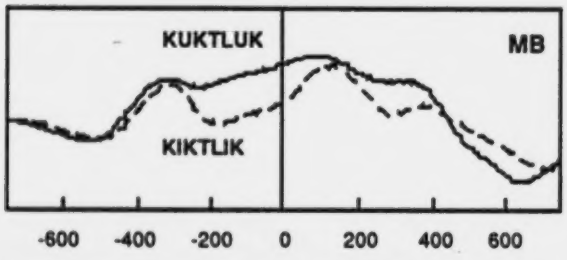


Figure 10. **Superimposed /kuktluk/ and /kiktlik/ averaged lower lip protrusion traces as produced by English subject MB. The vertical line is V_1 offset. Vertical and horizontal scales are as in *Figure 8*.**

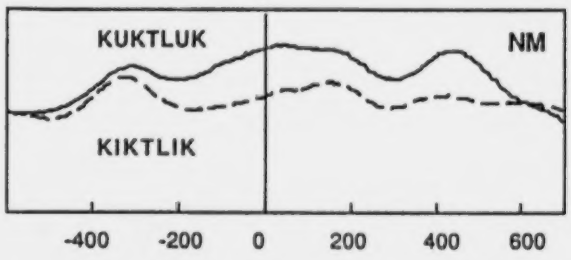


Figure 11. **Superimposed /kuktluk/ and /kiktlik/ averaged lower lip protrusion traces as produced by English subject NM. The vertical line is V_1 offset. Vertical and horizontal scales are as in *Figure 9*.**

These observations suggest that the central peak of the lower lip trace for /kuktluk/ may be largely due to protrusion for one or more of the intervocalic consonants. The protrusion in MB's upper lip trace may also reflect movement for consonants.[5] The implication of these observations is that protrusion movement during the consonants in /kuktluk/ is independent of its rounded vowel context. It is interesting, in this context, that both EMG OOS and OOI signals for /kuktluk/ are less strong during the consonant interval than during the rounded vowels. It seems likely that some or all of the consonant interval protrusion for these signals is due to jaw activity, perhaps as a consequence of jaw raising for the consonantal occlusions.[6]

Summary

The basic question behind the experiment was whether English and Turkish speakers would show the same articulatory patterns when producing similar words with rounded vowels separated by rounding-neutral consonants. The data presented here indicate that they do not. Rather than showing the consistent trough-like movement and EMG patterns exhibited by English speakers AE and AF, and reported in the literature for speakers of English, Swedish, Spanish and French, Turkish speakers show a consistent plateau-like pattern of movement and a unimodal pattern of EMG activity (with the possible exception of the ULX signal for subject EG). Equally, the Turkish subjects' patterns of movement and EMG contrast with the multi-peaked movement pattern and trough-like EMG patterns of English speakers MB and NM. Additionally, the latter two groups differ in the degree of consonant-related protrusion seen in i-i utterances.

The look-ahead model, as modified by Engstrand (1981), might account for these data in the following way: (1) for English speakers such as AE and AF, full lip protrusion (i.e., to the degree found in rounded vowels) is prohibited during one or more of the intervocalic consonants used in this study (see footnote 5); (2) for English speakers such as MB and NM, full lip protrusion for one or more consonants is required; (3) for Turkish speakers lip protrusion is compatible with but not required during these consonants, so that the degree of lip protrusion seen is dictated by feature spreading from the segmental context. Thus, for the English speakers, consonants must have some phonetic feature specification associated withprotrusion, although this may be either plus or minus, while for the Turkish speakers consonants are allowed to have neutral specification for this feature. It should be noted that for this version of the look-ahead theory, because the context-independence of gestures is not a theme, there is no straightforward prediction of relationship between the u-u and i-i word data. It is possible to say, for instance, that for speakers such as AE and AF lessened protrusion on consonants is a reaction to a strongly protruded environment, and the behavior of the same consonants in an unrounded environment is irrelevant.

For the coproduction model, in which articulatory output trajectories are the result of combining sequences of relatively stable, independently organized gestures, data from other contexts such as the i-i words becomes more important. In this model, the fact that the central peak in the u-u word movement traces for English subjects MB and NM has a counterpart in the i-i word traces is particularly relevant, as it suggests that the consonant-related peak in the u-u word traces may be independent of the gestures for the flanking vowels. Similarly, the relative lack of movement in the i-i word traces for speakers AE and AF suggests that the trough patterns in their u-u word traces result from combining overlapping vowel gestures with a small or non-existent consonant gesture. For the Turkish data, on the other hand, the lack of movement associated with the consonant(s) in the i-i word traces, together with the lack of a trough pattern in the u-u word traces, means that a different explanation is called for. According to the coproduction model, there are several possibilities. First, gestures for rounded vowels in Turkish (in contrast to those for English) may simply combine so as to produce a plateau pattern. This could happen, for instance, if Turkish gestures were larger or if the gesture-to-gesture interval were shorter, such that their overlap results in little or no trough. Alternatively, Turkish may have a different algorithm for combining gestures. Finally, the peculiar phonological properties of vowel harmony may result in successive rounded segments being associated with the same protrusion gesture.

The differences seen here between Turkish and English are also interesting in terms of the other theories mentioned above. For instance, if the trough in English is assumed to be a marker of syllable boundary, then the plateau pattern in Turkish may be taken to indicate that Turkish does not mark syllable boundaries in this way. Further, Turkish vowels such as /u/ are

(reportedly) not diphthongized, so that the lack of a trough in Turkish is compatible with a diphthongal account of the trough in English. Note, however, that for these theories the lack of a trough for English subjects MB and NM is somewhat problematic. It is necessary to assume either that the explanation does not apply to all English speakers or that the specification of protrusion for the intervening consonants obscures, in some fashion, the marking of syllable boundaries or the pattern of diphthongization.

Part II

Given the data reported here, it is not possible to test either the look-ahead, the syllable marker, or the diphthongization theories further. However, the (phonetic) context-free provision of the coproduction theory makes it amenable to testing based on articulatory behavior in different phonetic contexts. In essence, the logic is as follows: if articulator trajectories over several segments reflect the combination of gestures for each of the segments, then it should be possible to deduce the basic shape of each gesture from its behavior in different contexts. It should also be possible to synthesize articulatory contours by combining their elements.

Accordingly, this section of the paper describes a series of tests based on the context-free provision of the coproduction model. In the first test, the consonant-related protrusion gestures seen for English subjects MB and NM in ɪ-ɪ words are subtracted from corresponding protrusion traces for ʊ-ʊ words. Success is a function of correspondence, for the same speaker, between subtracted traces and other ʊ-ʊ word traces, such as EMG traces, with no suggestion of consonant interval protrusion. In other words, because the coproduction interpretation of inconsistencies between upper lip, lower lip, and EMG signals for these speakers involves the presence of an independent consonant-related protrusion gesture, removing the additional gesture should resolve the inconsistencies. In the second test, it is assumed that, if the vowel- and consonant-related gestures seen in the corpus are independently organized, then it should be possible to construct a viable ʊ-ʊ word from elements in ɪ-ʊ and ʊ-ɪ words. Thus, the original protrusion traces from ɪ-ʊ and u-i words are added together and the result compared to original ʊ-ʊ word traces. Success here is a function of degree of correspondence between original ʊ-ʊ word signals and the synthesized versions. The use of subtraction and addition for gesture combination is based on data reported by Löfqvist (1989), Saltzman, Rubin, Goldstein and Browman (1987); Saltzman and Munhall (1989).

Both tests require similar intersegment timing of consonant and vowel gestures in the ɪ-ɪ, ʊ-ʊ and mixed-vowel words. Although explicit measures of gestural timing were not made, mean intervocalic consonant intervals (from acoustic offset of V_1 to acoustic onset of V_2) among one-, two- and three-consonant words varied by less than 35 ms for any English or Turkish subject. This was taken as evidence that speech rate and gesture phasing were similar enough for corresponding gestures to be equated.

Subtraction Test

Those ɪ-ɪ word signals showing protrusion in the consonant interval consisted of upper and lower lip movement traces for subject MB and lower lip traces for subject NM. For the first test (henceforth called the Subtraction Test), these traces were subtracted, point by point, from corresponding u-u word movement traces. The theory behind this procedure was that the underlying movement during the consonant interval, i.e., the portion of movement associated with the vowel gestures, would be the same for both ɪ-ɪ and ʊ-ʊ words.[7]

Figures 12 and 13 show the results of subtracting averaged /kiktlik/ from averaged /kuktluk/ movement traces, superimposed on original /kuktluk/ movement traces for these subjects. For comparison purposes, Figure 13 also shows the results of subtracting NM's averaged upper lip /kiktlik/ movement trace, which showed no sign of protrusion during the consonant interval, from her averaged upper lip /kuktluk/ trace.

All four subtracted traces in Figure 10 show a trough pattern. This is to be expected for the upper lip trace of subject NM, since her original /kuktluk/ traces showed a trough and her /kiktlik/ trace is essentially flat. It is striking, however, that the trough patterns for both subjects' lower lip traces, and for MB's upper lip trace, correspond more neatly to these subjects' EMG trough patterns (seen in Figures 8 and 9) than did the original ʊ-ʊ traces. Further, NM's subtracted lower lip trace is nearly identical to both her subtracted and original upper lip trace.

This result supports the hypothesis that subjects NM and MB have separate vowel and consonant-related behavior for protrusion. Further, it suggests that their vowel-related protrusion behavior—presumably connected to articulatory instantiation of the vowel feature of

rounding—resembles that of other English speakers in being trough-like. The presence of lip protrusion during consonant articulation suggests, not a different articulatory organization, but an additional gesture overlapping with vowel-related gestures. At a more general level, this result can be taken as support for the coproduction model notion that gestures are independent entities and for the notion that gesture combination is approximately additive.

Addition Test

For the second test (henceforth known as the Addition Test), averaged upper and lower lip movement ɪ-ʊ and ʊ-ɪ word traces were added together for each English and Turkish subject. Because the result of adding ɪ-ʊ and ʊ-ɪ word traces is theoretically equal to the result of adding ʊ-ʊ and ɪ-ɪ word traces, the averaged ɪ-ɪ word traces were then subtracted from each added trace[8] to produce a "constructed" ʊ-ʊ trace. Figures 14 - 21 show the results of this procedure for lower lip data from /kuktluk/, /kiktlik/, /kuktlik/ and /kiktluk/ (upper lip data are substantially the same). The top panels show averaged /kuktlik/ and /kiktluk/. The traces resulting from adding these and subtracting /kiktlik/ (henceforth known as constructed traces) are shown in the bottom panels together with superimposed original ʊ-ʊ traces.

As these figures show, the constructed traces paralleled the original traces quite closely for three out of four English subjects, and for two out of the four Turkish subjects. For English subjects MB and AE, in particular, the traces parallel one another quite closely. For Turkish subject CK the principal difference is the slightly lower amplitude of the constructed trace.[9] For English subject NM, differences are also minimal. For Turkish subject EG, differences are intensification of a slight "bump" existing in the original trace plus a slightly lowered amplitude of movement during the final /u/ vowel. For the remaining subjects, however, differences are more serious. English subject AF's constructed trace shows a single broad peak rather than a trough as in the original trace. In contrast, Turkish subject IB's constructed trace shows a trough rather than a plateau as in the original trace. Turkish subject AT's constructed trace, while paralleling the original trace during the final vowel, has a generally different shape from the plateau pattern of the original trace.

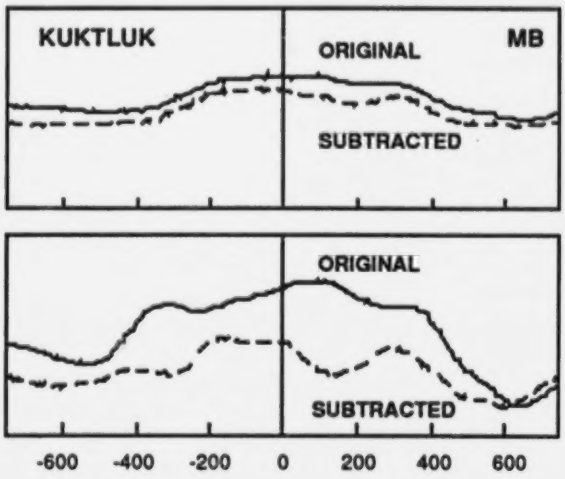


Figure 12. Original averaged /kuktluk/ protrusion traces (solid lines) with superimposed trace achieved by subtracting original averaged /kiktlik/ trace from original averaged /kuktluk/ trace (dashed line), for English subject MB. Upper panel shows original and subtracted traces for the upper lip, lower panel shows the same for the lower lip. The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 8.*

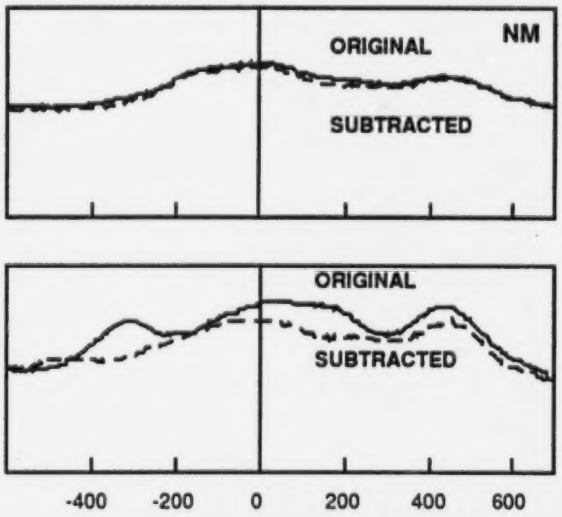


Figure 13. Original averaged /kuktluk/ protrusion traces (solid line) with superimposed trace made by subtracting original averaged /kiktlik/ trace from original averaged /kuktluk/ trace (dashed line), for English subject NM. Upper panel shows original and subtracted traces for the upper lip, lower panel shows the same for the lower lip. The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 9.*

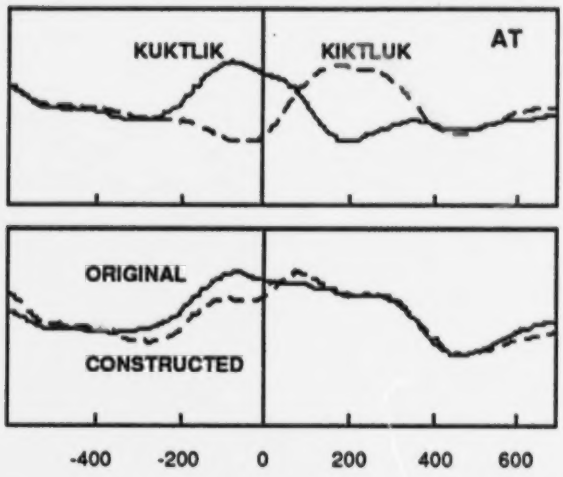


Figure 14. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by Turkish subject AT. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 2*.**

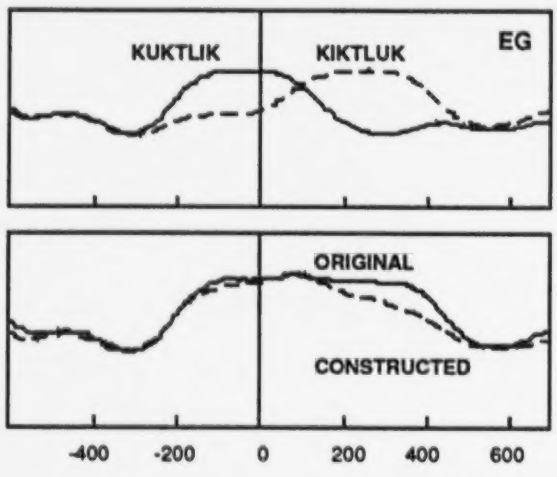


Figure 16. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by Turkish subject EG. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 4*.**

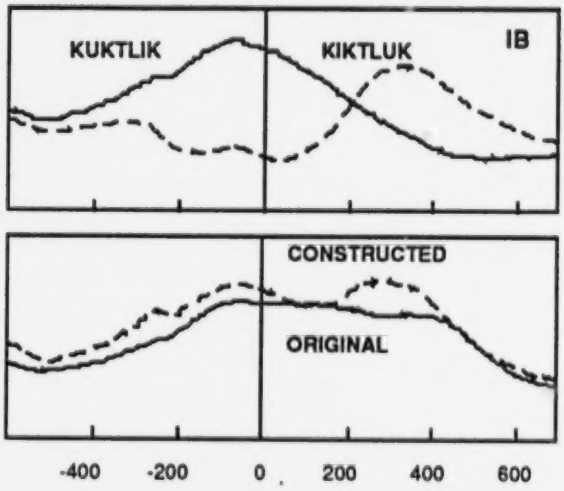


Figure 15. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by Turkish subject IB. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 3*.**

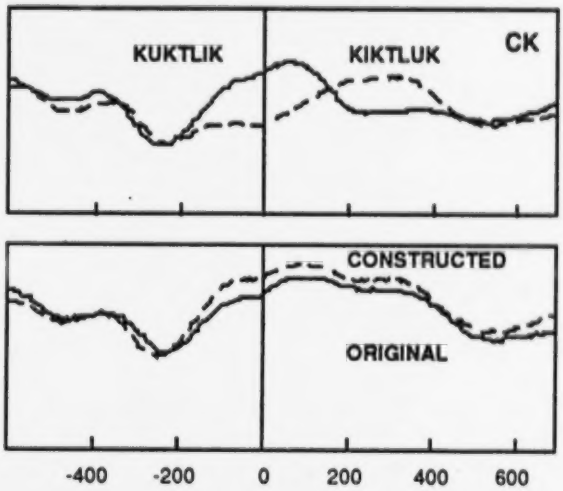


Figure 17. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by Turkish subject CK. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 5*.**

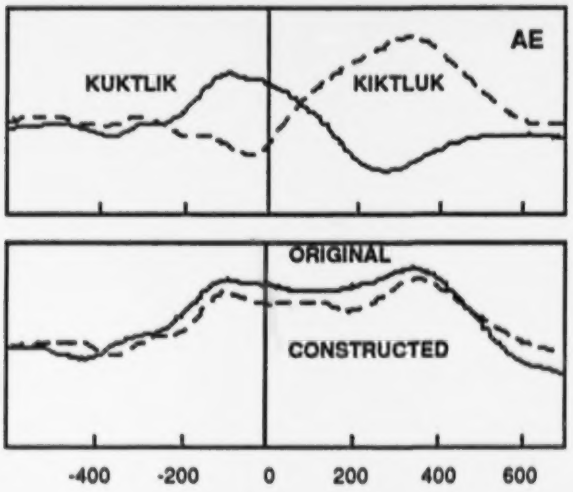


Figure 18. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by English subject AE. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 6*.**

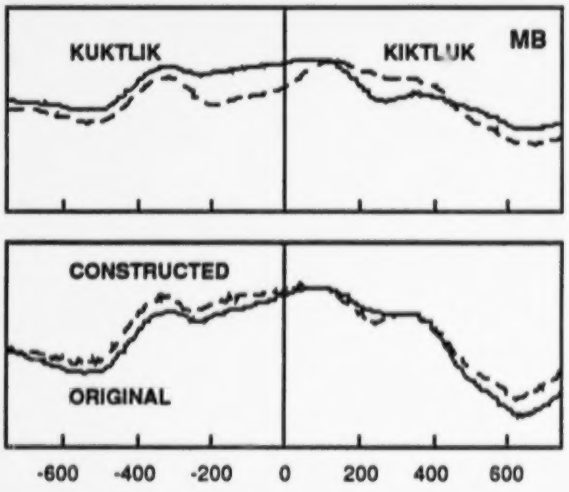


Figure 20. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by English subject MB. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 8*.**

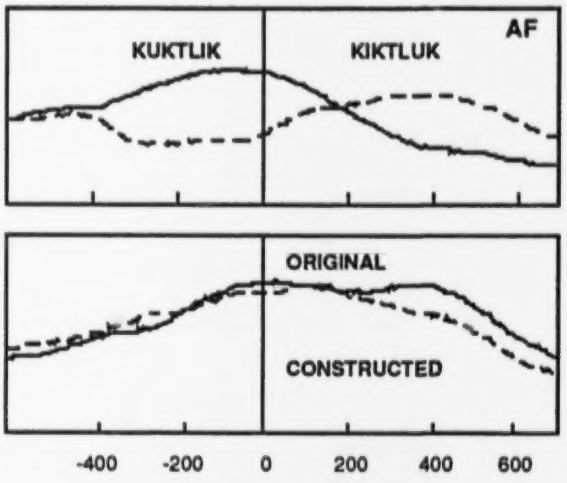


Figure 19. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by English subject AF. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 7*.**

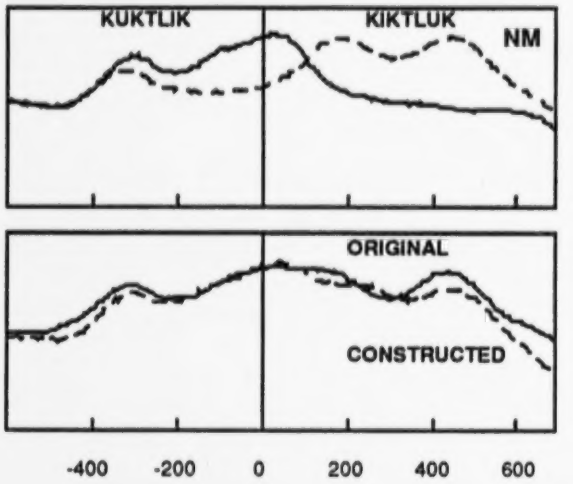


Figure 21. **Upper panel shows averaged lower lip protrusion traces for /kuktlik/ and /kiktluk/ as produced by English subject NM. Lower panel shows original averaged protrusion trace for /kuktluk/ (solid line) with superimposed trace constructed by adding averaged traces for /kuktlik/ and /kiktluk/ and subtracting averaged trace for /kiktlik/ (dashed line). The vertical line is the V_1 offset. Vertical and horizontal scales are as in *Figure 9*.**

DISCUSSION

The results of the Subtraction Test constitute relatively strong evidence for the generalization that English speakers produce troughs for words such as /kuktluk/, and for the notion that gestures are independent entities whose trajectories combine when overlapped in time. The subtraction test results also suggest that additivity is at least a reasonable approximation of the way that gestures combine for these articulators and these segments.

The results of the Addition Test are less clear. While the predicted and actual trajectories were close for some subjects, for other subjects they were qualitatively different. While the results were slightly better for English subjects than for Turkish subjects, the distinction between three subjects out of four (for English) vs. two subjects out of four, or even one out of four (for Turkish), is hardly great enough to warrant concluding the two languages are different. It is also not clear how to interpret a lack of correspondence between constructed and original traces; for instance, the assumption of similar conditions of speech rate, stress and gesture phasing between averaged ɪ-ʊ, ʊ-ɪ, ʊ-ʊ and ɪ-ɪ words may not be accurate. The fact that in Turkish ɪ-ʊ and ʊ-ɪ words are non-harmonic is also relevant. It is possible, for instance, that /u/ and /i/ vowels in Turkish words are always produced with independently organized gestures, but that these gestures are different in harmonic and non-harmonic words. A fuller discussion of these issues can be found in Boyce (1988).

General Discussion

Overall, the results of this study suggest there is something very different in the way English and Turkish speakers organize articulation, at least in the way they use lip protrusion for rounded segments. The simplest index of this difference is the plateau pattern of protrusion evinced by the Turkish speakers, which contrasts with the English patterns found here, and with the trough patterns reported in the literature to date. The results for English subjects, and in particular the results of the subtraction test for subjects MB and NM, confirm that the underlying articulatory strategy for ʊ-ʊ words in English follows a trough pattern.

With regard to the competing coproduction and look-ahead types of models, interpretation of these results is both straightforward and complex. The straightforward interpretation is as follows. Since the coproduction model predicts a trough pattern in ʊ-ʊ words, and English shows a trough pattern, then English speakers employ a coproduction articulatory strategy. Since the look-ahead model predicts a plateau pattern in ʊ-ʊ words, and Turkish shows a plateau pattern, then Turkish speakers employ a look-ahead strategy. Thus, English and Turkish have different articulatory strategies.

This interpretation gains strength from the fact that, for each model, explaining the patterns of both English and Turkish requires an additional mechanism. To explain the English trough pattern the look-ahead model must posit additional effects such as syllable-boundary marking, diphthongization, or consonant-specific unrounding in a rounded context. Similarly, to explain the Turkish plateau pattern the coproduction model must posit an unknown effect that causes ɪ-ʊ and ʊ-ɪ vowel gestures to differ from those in ʊ-ʊ words, or an unknown principle of gesture combination, or a loosening of the notion that gestures may be associated with only one segment. While any of these posited effects may ultimately prove to be valid, their status at this stage of investigation appears to be weak.

The complexity of this interpretation lies in the conclusion that different languages may employ different articulatory strategies. In some sense, this is to be expected, since the combination of phonology, lexicon and syntax in different languages may impose entirely different challenges to articulatory efficiency. In fact, the hypothesis behind this comparison of Turkish and English was the notion that, in contrast to English, Turkish provides ideal conditions for articulatory look-ahead. At the same time, human beings presumably come to the task of language acquisition with the same tools and talents. The finding that current models of coarticulation are insufficient to account for language diversity indicates that we have not yet penetrated to the universal level in the way we think about speech production. Further research, and in particular more cross-linguistic research, is needed in order to close this gap.

REFERENCES

Beckman, M., & Shoji, A. (1984). Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese. *Phonetica, 41*, 61-71.

Bell-Berti, F., & Harris, K. S. (1974). More on the motor organization of speech gestures. Haskins Laboratories *Status Report on Speech Research, SR-37/38*, 73-77.

Bell-Berti, F., & Harris, K. S. (1979). Anticipatory coarticulation: some implications from a study of lip rounding. *Journal of the Acoustical Society of America, 65*, 1268-1270.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Bell-Berti, F., & Harris, K. S. (1982). Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America, 71*, 449-454.

Benguerel, A.-P., & Cowan, H. (1974). Coarticulation of upper lip protrusion in French. *Phonetica, 30*, 41-55.

Blair, C., & Smith, A. (1986). EMG recording in human lip muscles: Can single muscles be isolated? *Journal of Speech and Hearing Research, 29*, 256-266.

Boyce, S. (1978). *Accent or accident?: The acoustical correlates of word-level prominence in Turkish.* Unpublished Batchelor of Arts thesis, Harvard University (Linguistics Department).

Boyce, S. E. (1988). *The influence of phonological structure on articulatory organization in Turkish and in English: Vowel harmony and coarticulation.* Unpublished doctoral dissertation, Yale University.

Brown, G. (1981). Consonant rounding in British English: The status of phonetic descriptions as historical data. In R. E. Asher & E. J. A. Henderson (Eds.), *Towards a history of phonetics.* Edinburgh: Edinburgh University Press.

Clements, G. N., & Sezer, E. (1982). Vowel and consonant disharmony in Turkish. In H. van der Hulst & N. Smith (Eds.), *The structure of phonological representations.* Foris, Dordrecht.

Cohn, A. C. (1988). Phonetic evidence for configuration constraints. Presented at the 19th Meeting of the New England Linguistic Society (NELS), November, 1988.

Daniloff, R. G., & Moll, K. L. (1968). Coarticulation of lip-rounding. *Journal of Speech Hearing Research, 11*, 707-721.

Engstrand, O. (1981). *Acoustic constraints of invariant input representation? An experimental study of selected articulatory movements and targets.* Reports from Uppsala University, Department of Linguistics, *7*, 67-94.

Fowler, C. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics, 8*, 113-133.

Gay, T. J. (1978). Articulatory units: Segments or syllables? In A. Bell & J. B. Hooper (Eds.), *Syllables and segments.* Amsterdam: North-Holland Press.

Gelfer, C. E., Bell-Berti, F., & Harris, K. S. (1989). Determining the extent of coarticulation: Effects of experimental design. *Journal of the Acoustical Society of America, 86*, 2443-2445.

Harris, K. S., & Bell-Berti, F. (1984). On consonants and syllable boundaries. In L. Raphael, C. Raphael, & M. Valdovinos (Eds.), *Language and cognition.* New York: Plenum Publishing.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Keating, P. (1985). CV phonology, experimental phonetics and coarticulation. *UCLA Working Papers in Phonetics, 62*, 1-13.

Keating, P. (1988). Underspecification in phonetics. *Phonology, 5*, 275-292.

Kozhevnikov, V. A., & Chistovich, L. A. (1966). Rech, artikulatsiya, i vospriyatiye, [Speech: Articulation and perception] (originally published 1965). Washington, DC: Joint Publications Research Service.

Ladefoged, P. (1975). *A course in phonetics.* New York: Harcourt Brace Jovanovich.

Leidner, D. R. (1973). *An electromyographic and acoustic study of American English liquids.* Unpublished doctoral dissertation, University of Connecticut, Storrs.

Lewis, G. L. (1967). *Turkish grammar.* Oxford: Clarendon Press.

Liberman, A. M., & Studdert-Kennedy, M. (1977). Phonetic perception. In R. Held, H. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of sensory physiology, Vol. VIII: Perception.* Heidelberg: Springer-Verlag.

Löfqvist, A. (1989). Speech as audible gestures. Paper presented at the NATO Advanced Study Institute Conference on Speech Production and Speech Modeling, Bonas, France.

Lubker, J. (1981). Temporal aspects of speech production: Anticipatory labial coarticulation. *Phonetica, 38*, 51-65.

Lubker, J., & Gay, T. (1982). Anticipatory labial coarticulation: Experimental, biological, and linguistic variables. *Journal of the Acoustical Society of America, 71*, 437-447.

Magen, H. (1984). Vowel-to vowel coarticulation in English and Japanese. *Journal of the Acoustical Society of America, 75*, S41.

Manuel, S. Y. (1990). The role of output constraints in vowel-to-vowel coarticulation. Submitted to *Journal of the Acoustical Society of America,*

Martin, J. G., & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance 8*, 473-488.

McAllister, J. (1978). Temporal asymmetry in labial coarticulation. *Papers from the Institute of Linguistics, 35*, 1-29, (University of Stockholm, Stockholm).

Ohala, J. J. (1981). The listener as a source of sound change. In C. Masek, R. Hendrick & M. F. Miller (Eds.), Papers from the parasession on language and behavior. Chicago: Chicago Linguistic Society.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151-168.

Perkell, J. S. (1986). Coarticulation strategies: preliminary implications of a detailed analysis of lower lip protrusion movements. *Speech Communication, 5*, 47-68.

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech, 28*, 97-114.

Saltzman, E. L., & K. G. Munhall. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.

Saltzman, E., Rubin, P. E., Goldstein, L., & Browman, C. P. (1987). Task-dynamic modeling of inter-articulator coordination. *Journal of the Acoustical Society of America, 82*, S15.

Sussman, H. M., & Westbury, J. R. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research, 46*, 16-24.

Tuller, B., Kelso, J. A. S., & Harris, K. S. (1982). Interarticulator phasing as an index of temporal regularity in speech. *Journal of Experimental Psychology: Human Perception and Performance 8*, 460-472.

FOOTNOTES

*To appear in *Journal of the Acoustical Society of America.*

†Speech Communication Group, Research Laboratory of Electronics, Massachusetts Institute of Technology.

[1] The models tend to differ in the level at which compatibility is assessed. In Henke's program, the limitation was explicitly defined on an articulatory basis. Other investigators (Cohn, 1988; Keating, 1988) have postulated that coarticulation spreads by reference to feature specification at the phonological level.

[2] Rotational head movement could theoretically affect the results in two ways. First, the LED's might have been moved into a more peripheral area of the focus field where tracking is less accurate. Subjects were constantly monitored against this possibility during the course of the experiment, and videotaped experiments were checked post-hoc. Second, rotation of the head changes the relationship between the vertical and horizontal axes of the LED tracking system (the X and Y coordinates) and the subjects' sagittal midline. Thus, less or more of the subjects' anterior-posterior movement relative to the midline may be detected. Note, however, that in all cases

where a subjects' amplitude of movement changed over the course of the experiment this change was mirrored in the corresponding EMG signal, suggesting that baseline change was not a significant factor.

[3] Interestingly, for subjects IB and CK, the EMG pattern for u-u words resembled that for u-ɨ words. The pattern for ɨ-u words showed a strong peak associated with V_2. For subject EG, on the other hand, the pattern for /kuktluk/ resembled that for /kiktluk/, while the patterns for shorter words /kukuk/, /kuluk/, etc. resembled those for /kukik/, /kulik/, etc. EMG traces for these words are reported in Boyce (1988).

[4] To some extent this difference between OOS and OOI signals is a consequence of averaging, as token traces for the two signals showed differing proportions of double- and triple-peaked patterns.

[5] Perusal of the traces for /kitik/, /kilik/ and /kikik/ suggest that NM shows some lower lip protrusion for each intervocalic consonant. For MB the lower lip trace shows marked protrusion for /t/ and some protrusion for /l/, while the upper lip shows protrusion only for /l/. Each of these protrusion peaks matches the consonant-interval peak of the corresponding u-u word.

[6] Other comparisons of JX, JY, LLX and LLY signals, as well as observations of the videotapes, suggest that these speakers use forward movement of the jaw (both rotational and translational) as well as lip movement to produce protrusion during rounded vowels. This topic is discussed further in Boyce (1988).

[7] One intractable problem with this procedure is that /i/ and /u/ vowels also should have characteristic patterns. Thus, if retraction for /ɨ/ vowels is present it will be subtracted from protrusion for /u/ vowels at the same time the consonant protrusion is subtracted. In these data, the relative magnitude of protrusion dwarfed that of retraction. Thus, it was assumed that the effects of subtracting the one outweighed the effects of the other.

[8] As in the first test, some residue of possible /ɨ/ vowel-associated movement remains in these constructed traces.

[9] Note that the theory of independent gestures does not require that all gestures have identical amplitude or be produced with identical force. Such a requirement would leave no room for the effect of fatigue or for prosodic variables such as stress and syllable position. It is a common observation, for instance, that EMG or movement signals for the same word may show less amplitude at later stages during the same experiment. In this study, the fact that ɨ-u, u-u, u-ɨ and ɨ-ɨ utterances were blocked separately may have caused some differences in overall amplitude among them.

APPENDIX

Turkish has eight vowels /i ɨ a e o ø u y/ and thus (like Swedish but unlike English) has vowels which contrast only in rounding. The consonants /t/, /k/ and /l/ are non-labial and phonemically unrounded in both languages (Ladefoged, 1975; Lewis, 1967). English and Turkish have somewhat different patterns of allophonic variation for /k/ and /l/. In Turkish /k/and /l/ tend to be front or back according to the front/backness of the vowel of the same syllable (Clements & Sezer, 1982). In contrast, for most English dialects syllable-initial /l/ is front and syllable-final /l/ is back, i.e. velarized (Keating, 1985; Ladefoged, 1975), while /k/ varies primarily in syllable-initial position, becoming front before front vowels and back before back vowels. (Although it is sometimes referred to as "consonant harmony," the Turkish rule for /l/ and /k/ is distinct from that for front/back harmony in vowels.) Neither /l/ nor /k/ participate in roundness harmony. The sequence /ktl/ is rare in both languages, but exists, c.f. English *tactless* and Turkish /paktlar/ 'pacts.' Neither language allows the initial cluster /tl/; therefore, /kuktluk/ would have the syllable structure /kukt-luk/ in both languages.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 19-37

Long Range Coarticulatory Effects for Tongue Dorsum Contact in VCVCV Sequences*

Daniel Recasens†

The goal of this paper is to gather accurate information about the temporal and spatial properties of tongue dorsum movement in running speech. Electropalatographic and acoustical data were collected to measure lingual coarticulation over time. Coarticulatory effects were measured along VC[ə]CV utterances for articulations differing in the degree of tongue dorsum contact, namely, for vowels [i] vs. [a] and for consonants [ʃ] vs. [t]; the contextual phonemes were all possible combinations of those same consonants and vowels. Results show contrasting mechanisms for anticipatory and carryover coarticulation. Anticipatory effects appear to be more tightly controlled than carryover effects presumably because of phonemic preplanning; accordingly, gestural antagonism in the contextual phonemes affects the two coarticulatory types differently. The relevance of these data with respect to theories of coarticulation and speech production modeling is discussed.

I. INTRODUCTION

A large body of experimental evidence in the phonetics literature shows that the articulatory gestures for successive phonemic units are coarticulated in running speech and thus overlap in time. Many studies of coarticulation aim at reaching some understanding about the mechanisms of phonemic realization in speech production. A relevant goal of this research is to separate articulatory events resulting from motor programming strategies, from others related to the mechanical and physical properties of the speech production system. This paper approaches the nature of these two components through an analysis of tongue dorsum coarticulation over time, under the theoretical assumption that the phonemic gestures are programmed to exhibit certain temporal and spatial patterns, and that those patterns prevail to a large extent across changes in phonemic context, speech rate and speaker (Harris, 1984; MacNeilage & DeClerk, 1969).

The research to be reported here is relevant to a number of significant issues.

A. Coarticulation at a distance

One of the issues of interest in coarticulation studies is the temporal domain of gestural activity for a given phonemic unit. Early research suggested that jaw (Gay, 1977; Sussman et al., 1973) and lingual activity (see references in section I.C) did not extend more than one or two segments beyond the target phoneme; more recent acoustic evidence shows however that such coarticulatory effects may last for three or four phonemes. Thus, long range temporal effects have been found for American English in [VləlV] (V1-to-V3 effects; Huffman, 1986) and in [VbəbV] (V3-to-V1 effects; Magen, 1989) sequences. The present paper intends to replicate these findings through an analysis of coarticulation at a distance in VCVCV utterances. Analogously to Huffman's and Magen's studies, V2 was kept constant as [ə] to facilitate long range effects over time since the schwa is highly sensitive to coarticulation from the adjacent segments (Recasens, 1985). Another goal of this research is to find out whether long range coarticulatory effects occur in other languages besides English.

I wish to thank C. A. Fowler, K. S. Harris and I. G. Mattingly for their valuable suggestions on particular aspects of this research. I would also like to thank R. Arens Krakow, E. Bateson, V. Gulisano and R. McGowan for technical help. This research was supported by NINCDS Grant NS-13617 to Haskins Laboratories, a postdoctoral fellowship from the Spain-U.S. Committee, research grant PB87-0774-C02-01 from the Spanish Government (DGICYT), and a fellowship from the Catalan Government (CIRIT Commission).

B. V-and C-dependent coarticulation

Evidence for transconsonantal V-to-V coarticulation, and for greater V-to-C effects than C-to-V effects (e.g., MacNeilage & DeClerk, 1969), suggests that consonantal gestures unspecified for tongue-dorsum activity (e.g., labials and dentoalveolars) are produced with reference to a continuous V-to-V cycle (Fowler, 1984; Öhman, 1966). The implication of this view is that the domain of consonant related coarticulation is more constrained than the domain of vowel related coarticulation. V-to-V effects ought to be larger than C-to-C effects since the phonemic string is organized according to underlying gestures of a diphthongal nature.

To test this theory I will study coarticulatory effects from vowels and consonants differing in degree of linguopalatal contact along VC[ə]CV sequences. Thus, C-dependent effects will be analyzed for the palatoalveolar consonant [ʃ] vs. the apicodental or apicoalveolar consonant [t], and V-dependent effects for the palatal vowel [i] vs. the non-palatal vowel [a]. If the hypothesis is correct, C-dependent effects ought not to extend beyond V2=[ə]. Notice however that since the schwa has been denied a vocal tract target of its own (Catford, 1977) the availability of C-to-C effects across [ə] would not invalidate a strong version of the theory. However, it would be critical for the theory if the temporal domain of the C-dependent effects exceeded that of the V-dependent effects.

C. Gestural antagonism

Theories of coarticulation agree that gestural activity corresponding to phonemic units does not take place across antagonistic gestures (see section I.D). The problem is that no clear formulation of what is meant by gestural antagonism is available in most cases (Bell-Berti & Harris, 1981; Fowler, 1984). It is proposed in this paper that the degree of tongue dorsum coarticulation is inversely related to the involvement of the dorsum of the tongue in making a closure or a constriction. Data for consonants and vowels in the literature offer some support for this hypothesis.

Anticipatory V-to-V effects for tongue dorsum activity in VCV utterances have been found by Butcher and Weiher (1976), and Alfonso and Baer (1982), but much less so or not at all by Carney and Moll (1971), Gay (1977), Parush et al. (1983), and Farnetani et al. (1985). The explanation underlying these contrasting findings lies partly in the articulatory characteristics of the intervocalic consonant. Thus, V-to-V effects occur mainly for consonants for which the dorsum of the tongue is not involved in the making of a closure or a constriction ([p]: Alfonso and Baer (1982); [t]: Butcher and Weiher (1976)). Velars, on the other hand, were found to block V-to-V effects to a much larger extent than labials and dentoalveolars in some of the previous studies. I have shown in this respect that the degree of transconsonantal V-to-V coarticulation is related inversely and monotonically to the degree of tongue-dorsum contact, for more palatal-like vs. more alveolar-like consonants (Recasens, 1984).

The articulatory characteristics of vowels also affect the extent of V-to-V coarticulation over time. Like palatal consonants, palatal vowels are more resistant to tongue dorsum coarticulation than vowels showing no constriction at the palatal place of articulation. Thus, smaller V-to-V effects on [i] than on [a] have been reported by Gay (1977), Butcher and Weiher (1976), and Recasens (1984).

This study will look into the extent to which coarticulation is blocked during the production of consonants and vowels involving different degrees of palatal contact, namely, [ʃ] and [t], and [i] and [a]. The hypothesis that coarticulation is inversely related to the degree of tongue dorsum contact for the contextual gestures will be tested across adjacent and distant phonetic segments; for that purpose, all possible combinations of consonants [ʃ] and [t], and vowels [i] and [a] in VC[ə]CV sequences will be submitted to experimental analysis.

D. Anticipatory vs. carryover coarticulation

Another issue of interest in the present study is the nature of the anticipatory and the carryover effects.

The sequencing of these two coarticulation types with respect to the target phoneme suggests that the anticipatory effects ought to reflect phonemic preplanning, and that the carryover effects ought to be mainly determined by the articulatory properties of the target and contextual phonemes. However, as shown in section I.C, the anticipation of tongue dorsum activity is not independent of the production characteristics of the preceding phonemes. Therefore, possible candidates for preplanning mechanisms would be those regularities in onset time of articulatory activity that can be shown to depend minimally on the articulatory properties of the preceding phonemic string.

Theories exist about what those regularities may be. They have been devised, however, to account mostly for velar and labial activity, and not so much for tongue movement (Sussman & Westbury, 1981). Moreover, they disagree as to whether anticipatory effects may affect an unlimited number of non antagonistic gestures (Henke, 1966), or whether they are locked in time to the target gesture provided that no articulatory conflict is involved (Bell-Berti & Harris, 1981). Further refinements need to be carried out on theories of coarticulation to accommodate differences in coarticulatory activity for different articulators as well as differences in gestural antagonism associated with the contextual phonemes.

Some findings reported in coarticulation studies are relevant with respect to the contrasting nature of anticipatory vs. carryover effects for tongue dorsum activity: (a) anticipatory effects are essentially temporal and their onset shows little variability, while carryover effects are essentially spatial and more variable (Gay, 1977; Parush et al., 1983); (b) carryover effects may be larger than anticipatory effects (Farnetani et al., 1985; Recasens, 1984) but also smaller (Butcher & Weiher, 1976). Acoustic evidence for larger carryover vs. anticipatory effects (Fowler, 1981; Huffman, 1986) and vice versa (Magen, 1989) is also available.

One can argue that while the findings reported in (a) result from the preplanned vs. mechanical nature of anticipatory vs. carryover coarticulation, respectively, those reported in (b) are dependent on differences in speech rate, speaker and phonemic context. Thus, higher speech rates are likely to bring about an increase in the amount of anticipatory coarticulation while reducing the degree of antagonism from the preceding gestures; on the other hand, slower speech rates presumably cause the mechanical constraints for the preceding gestures to overcome those anticipatory effects associated with the preplanning of the target phoneme. Therefore one might plausibly argue for a progressive increase of anticipatory vs. carryover effects as speech rate increases, and of carryover vs. anticipatory effects as speech rate decreases. Similarly, the relative salience of the anticipatory vs. carryover effects may depend on the degree of articulatory constraint associated with the gestures preceding and following the target phoneme: anticipatory effects are likely to prevail upon carryover effects when the phonetic segments following the target phoneme involve higher gestural requirements than those preceding it; on the other hand, carryover effects should be larger than anticipatory effects if the phonetic segments preceding the phonemic target are more constrained than those following it.

Within this framework, the present study is concerned with the nature and the relative salience of the anticipatory vs. carryover modes of coarticulation across segments showing different degrees of coarticulatory resistance.

II. METHOD

A. Articulatory analysis

Electropalatography (EPG) was used to analyze tongue dorsum contact over time. Electropalatographic data were collected for all possible VC[ə]CV combinations with C=[ʃ], [t] and V=[i], [a], and stress on the last syllable; all sequences were embedded in a "p ____ p" carrier environment.

Subjects read the list of sequences listed in Table 1. Two assumptions underlie the ordering of those sequences. In the first place, the degree of contextual interference with respect to coarticulatory effects ought to decrease as we move away from the target phoneme. Thus, e.g., V1-dependent effects ought to be more heavily influenced by the articulatory characteristics of C1 than by those of the phonetic segments placed at the other side of [ə] (i.e., C2 and V3). Secondly, phonetic segments involving more palatal contact (i.e., [ʃ] and [i]) ought to conflict with coarticulatory effects to a larger extent than those involving a lesser degree of palatal contact (i.e., [t] and [a]).

Table 1. *List of sequences used in the experiment. Stress was placed on the last syllable.*

1. piʃəʃip	5. piʃətip	9. pitəʃip	13. pitətip
2. paʃəʃip	6. paʃətip	10. patəʃip	14. patətip
3. piʃəʃap	7. piʃətap	11. pitəʃap	15. pitətap
4. paʃəʃap	8. paʃətap	12. patəʃap	16. patətap

A composite account of the two assumptions explains the particular ordering of the utterances in Table 1. Thus, the utterances in the table are ordered for decreasing degrees of resistance to carryover coarticulation associated with V1. In the first place, the offset of V1-dependent effects for [i] vs. [a] is expected to occur earlier for sequences 1 through 8 than for sequences 9 through 16 since C1=[ʃ] in the first set of utterances and C1=[t] in

the second. Within each set of sequences with a different C1, V1-dependent carryover effects ought to last for a shorter time when C2=[ʃ] than when C2=[t] (sequences 1 through 4 vs. 5 through 8, and 9 through 12 vs. 13 through 16). Finally, within each set of sequences showing a different C1 and a different C2, V1-dependent carryover effects ought to last less long when V3=[i] than when V3=[a] (sequences 1 and 2 vs. 3 and 4, 5 and 6 vs. 7 and 8, and so on).

A Catalan speaker from the Barcelona region (Re), and two American English speakers from New York City (Ra, Ba), repeated all utterances ten times with the artificial palate in place. Simultaneous recordings were made of the EPG and the acoustic signals. The American English speakers were asked to avoid flapping of phonemic /t/ and to make an alveolar stop instead; phonetic perception and inspection of the EPG data revealed that the consonant was always [t].

The electropalatographic system (Rion Electropalatograph model DP-01) has been described elsewhere (Recasens, 1984; Shibata et al., 1978). It is equipped with 63 electrodes arranged in five semicircular rows (see Figure 1) and allows displaying one pattern of contact every 15.6 ms. As shown in the figure, the electrodes can be grouped in four articulatory regions (i.e., alveolar, prepalatal, mediopalatal and postpalatal) for data analysis. The figure also shows that some electrodes are located along a median line; since electrodes on this line belong neither to the right nor to the left side of the palate, they were assigned to both sides when necessary. One can see that the number of electrodes decreases as rows become more central; thus, the outermost row has 8.5 electrodes on each side and the innermost row has 3.5.

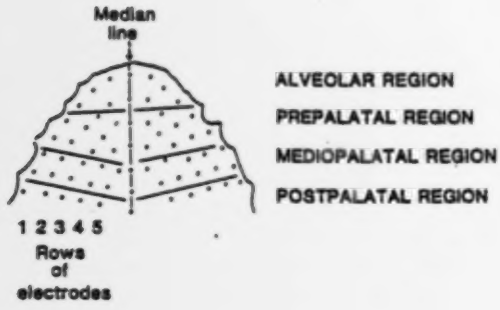


Figure 1. Electroplate.

B. Acoustical analysis

The acoustical data were digitized at a sampling rate of 10kHz, after pre-emphasis and low-pass filtering. An LPC program included in the ILS (Interactive Laboratory System) package was available for spectral analysis. F2 data were collected at the measurement points of interest and averaged across repetitions. F2 measurements were preferred to other measurements (e.g., F3 and/or F1) since, as shown in the literature, there is a good correlation between F2 and tongue placement for vowels (e.g., Alfonso & Baer, 1982; Fant, 1960).

C. Measurement points in time

The points in time selected for analysis are shown in Table 2 for the EPG and the F2 data.

Table 2. *Measurement points in time for the EPG data (above) and the F2 data (below).*

EPG data

1 2 3 4 5 6 7 8 9

1. V1 midp.
2. V1 off.
3. C1 midp.
4. V2 ons.
5. V2 midp.
6. V2 offs.
7. C2 midp.
8. V3 ons.
9. V3 midp.

F2 data

1 2 3 4 5

1. V1 midp.
2. V1 midp./offs.
3. V2 midp.
4. V3 ons./midp.
5. V3 midp.

Measurement points for the EPG data were labeled on the acoustic waveform except for the [t] midpoint (see Figure 2). This was due to the fact that, except for the [t] closure, the articulatory events of interest could not be located satisfactorily on the EPG record; thus, for example, it was not possible to select the frame corresponding to the period of maximum postalveolar constriction for [ʃ] (e.g., especially when adjacent to the vowel [i] as exemplified in Figure 2). Vowel onsets and offsets were labeled at onsets and offsets of voicing; moreover, those low amplitude pitch pulses occurring after the onset of lingual closure for [t], as determined visually on the acoustic waveform, were excluded from the

vocalic period in the measurement procedure. Vowel midpoints were labeled halfway between the vowel onsets and offsets. The [ʃ] midpoint was located at the midpoint between the voicing offset for the preceding vowel and the voicing onset for the following vowel. Acoustic waveforms for all utterances were labeled at the points in time indicated in Table 2 and the EPG frames corresponding to those acoustic labels were used for data analysis.

The consonantal midpoint for [t] was established at the closure midpoint. Since the closure period for this consonant is easy to determine on the EPG record it was decided to use an articulatory rather than an acoustic criterion in this case. Thus, the [t] midpoint was the only frame or the medial frame of several successive frames showing all electrodes lighted up on row 1; when this is the case, maximal contact at the front of the alveolar region is achieved.

Acoustical analysis (F2 measurements) was performed at different points along the vowels, as shown in Table 2. The main purpose of the acoustical analysis was to supplement the EPG data and, in particular, to obtain more information about coarticulatory effects at the midpoint of [a] since linguopalatal contact for this vowel at this particular point in time was practically absent for two of the three speakers (see Figure 5). The location of the vowel's midpoint is the same as that used in the EPG analysis. A single point along V2=[ɔ] was chosen for analysis (i.e., V2 midpoint). The V1 offset and the V3 onset labels were not placed at the voicing endpoints since the frequency of the F2 peak shows a good amount of variability at these points in time; instead, F2 measurements were taken at equidistant points between the V1 midpoint and the V1 offset (i.e., V1 midp./offs.), and between the V3 onset and the V3 midpoint (i.e., V3 ons./midp.).

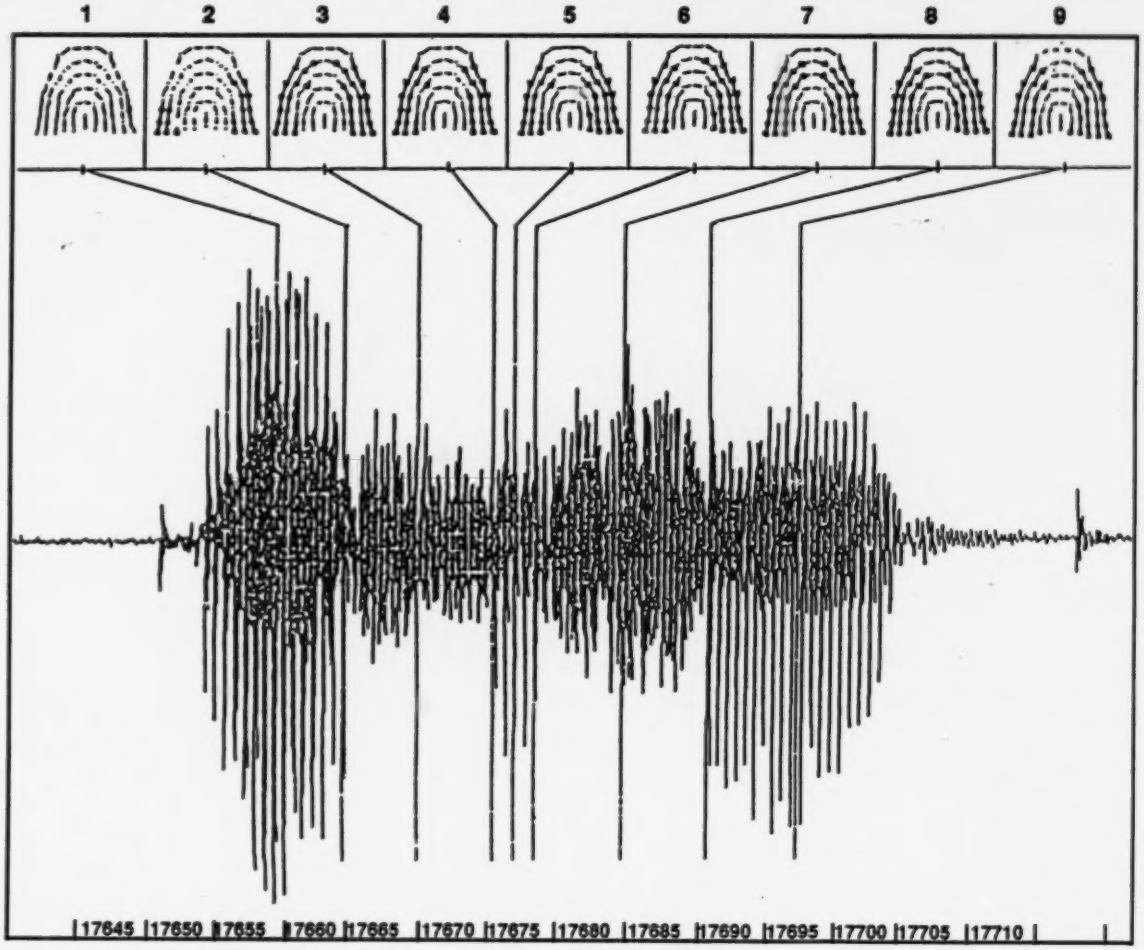


Figure 2. Exemplification of the line-up procedure between the acoustic labels and the EPG frames in the case of the utterance [paʃɔʃip] (speaker Ra). See Table 2 for measurement points in time 1 through 9.

D. Measurement of palatal contact

To achieve an accurate estimate of degree of tongue dorsum contact at the palatal regions, the measure of contact was the number of "on" electrodes in the innermost three or four rows of the artificial palate. The main reason for this choice was that differences between the degree of palatal contact for [i] vs. [a] and for [ʃ] vs. [t] are more salient at the center of the prepalate, mediopalate and postpalate, than at the sides of these articulatory regions (see section III.A). This criterion ensures that coarticulatory effects correspond to the most distinctive measure of contrast in degree of dorsopalatal contact, and are maximally preserved during the production of the surrounding phonemes.

EPG data was gathered for rows 3,4,5 for speakers Re and Ra, and for rows 2,3,4,5 for speaker Ba. This speaker-dependent contrast was due to the fact that linguopalatal contact was systematically more peripheral for speaker Ba than for speakers Re and Ra (see section III.A).

III. RESULTS

A. Linguopalatal contact and coarticulatory resistance for consonants and vowels

1. Consonants

Figure 3 shows patterns of linguopalatal contact averaged across repetitions at the midpoint of C2=[ʃ] and C2=[t] for all speakers. Linguopalatal configurations in the figure correspond to the stressed CV syllables in sequences 1, 4, 13 and 16 of Table 1.

In Figure 3, tongue contact takes place between the contour lines and the sides of the palate; the central area was left untouched by the tongue. Contour lines connect averages of linguopalatal contact between rows. The electrodes on or behind the contour line along a given row have been "on" 100% of contacts across repetitions. The degree of fronting for the contour line along the space between two adjacent electrodes on a given row is proportional to the contact average for the frontmost of the two electrodes; therefore the line lies closer to the frontmost electrode as the contact average for that electrode increases. Thus, in the case of the contour line for [ʃi] (speaker Ra), electrodes 1 through 6 on row 1 of the left side of the palate were lit up in 100% of occurrences (i.e., in all repetitions) and electrode 7 in 30% of occurrences (i.e., in three out of ten repetitions).

The patterns of contact for [ʃ] show that the consonant is produced with a central groove along the postalveolar and palatal regions while there is a wide contact area at both sides of the palate. As for [t], it is produced with a complete closure at the front of the alveolar region (all speakers) and at the postalveolar region (speaker Ra); overall, the degree of lateral contact is less than for [ʃ].

Figure 4 is another display of the same EPG frames row by row. The bars in the figure stand for the overall number of on electrodes on each row. As to the issue of concern here, namely, the degree of tongue dorsum contact at the center of the palatal regions (on rows 3,4 and 5; see Figure 1), a general trend is observed for [ʃ] to show more contact than [t]. However, while this is true before [a], it is not so much the case before [i]. Thus, the claim that [ʃ] shows more dorsal contact than [t] is always true in the adjacency of low vowels but not in the adjacency of high vowels. Moreover, as also shown in Figure 3, speaker Ba shows less central contact at the palatal regions than the other two speakers; thus, no contact at all was observed for [t] on rows 3, 4 and 5 . For that reason, the EPG data for this speaker were obtained by computing the number of on electrodes on rows 2, 3, 4 and 5 (see section II.D) at the expense of including a few electrodes located at the postalveolar region on row 2.

The EPG data on Figure 3 confirms the hypothesis that [ʃ] is less sensitive than [t] to coarticulatory effects in tongue dorsum activity because it involves more palatal contact. Thus, while the degree of contact at the palatal regions is highly similar for [ʃa] and [ʃi] (i.e., little or no tongue dorsum lowering occurs for [ʃ] before [a]), [ti] shows more tongue dorsum contact towards the median line than [ta] (i.e., during the production of [t], the tongue dorsum allows some vertical displacement as a function of the adjacent vowel). This trend is highly consistent for all three speakers. It should be pointed out that manner of articulation requirements may contribute to the absence of coarticulatory effects in tongue dorsum activity for [ʃ]; as shown in the literature (McCutcheon et al., 1980; Wolf et al., 1976), the formation of the medial groove for fricatives involves a high degree of articulatory precision.

In summary, adjacent vowels cause more variability in tongue dorsum contact for [ʃ] than for [t] in accordance with the articulatory characteristics of these consonants.

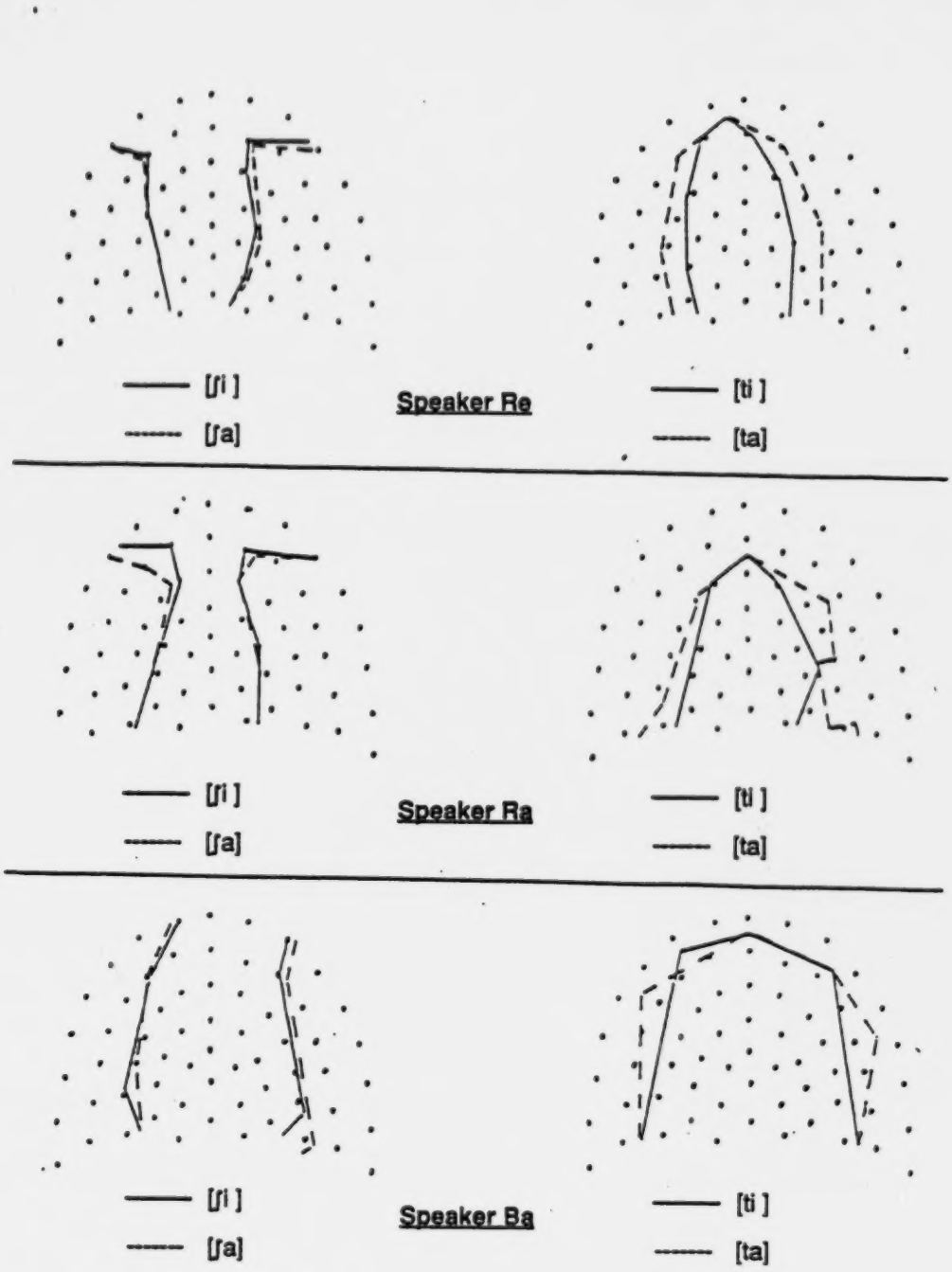


Figure 3. **Linguopalatal patterns at the midpoint of C2=[ʃ] and C2=[t] as a function of following [i] and [a] for speakers Re, Ra and Ba.**

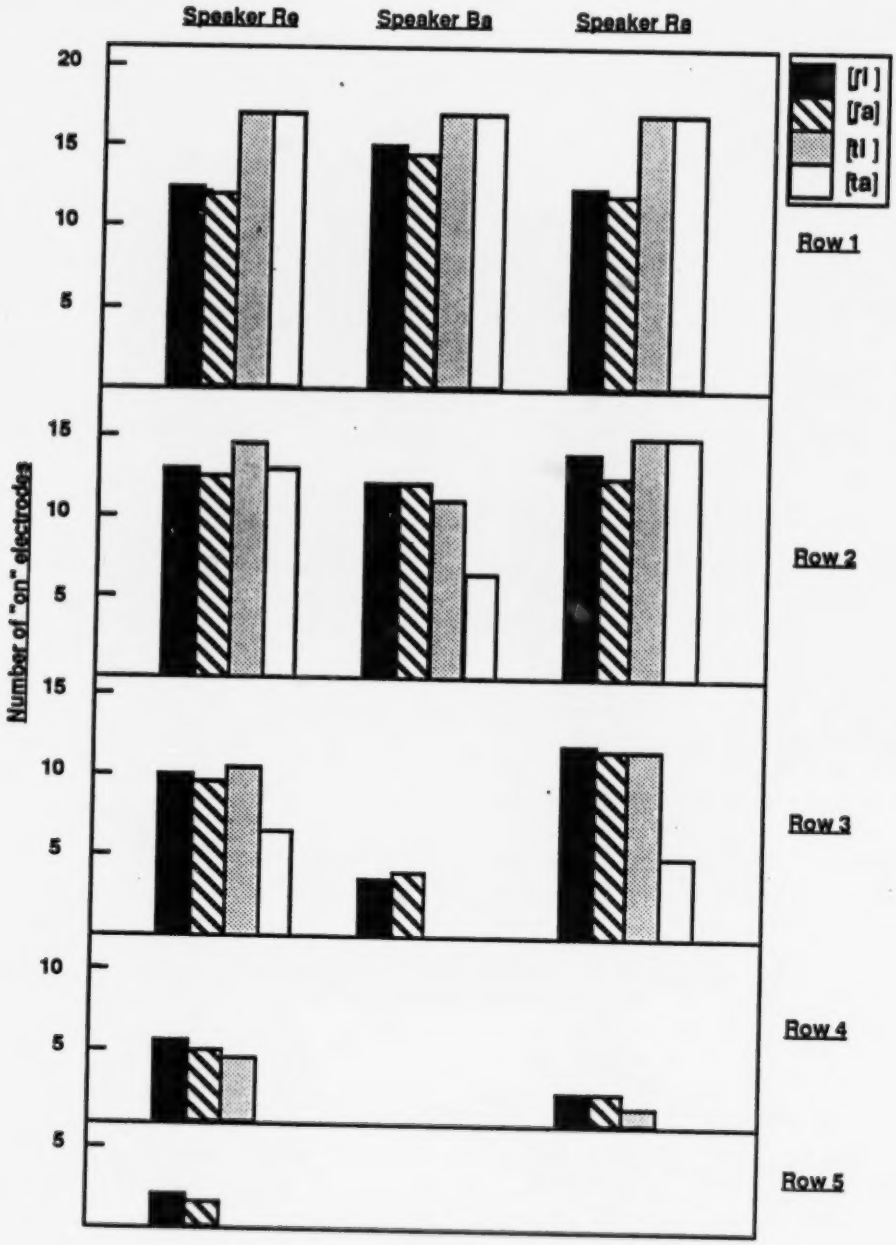


Figure 4. Number of "on" electrodes row by row at the midpoint of C2=[ʃ] and C2=[t] as a function of following [i] vs. [a] for speakers Re, Ra and Ba.

2. Vowels

Figure 5 shows patterns of linguopalatal contact averaged across repetitions at the midpoint of V3=[i] and V3=[a] for the three speakers. As for [ʃ] and [t] (Figure 3), only data for the stressed CV syllables in the symmetrical sequences are given.

Linguopalatal patterns show a larger area of contact at the center of the palatal regions and more tongue fronting for [i] than for [a]. The absence of linguopalatal contact for [a] in the case of speakers Ba and Ra shows that the vowel is lower in their New York City dialect than in Catalan.

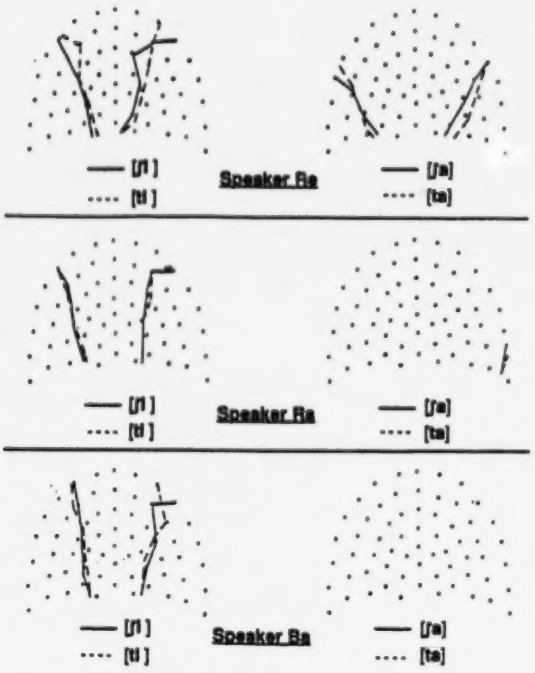


Figure 5. **Linguopalatal patterns at the midpoint of V3=[i] and V3=[a] as a function of preceding C2=[ʃ] vs. [t] for speakers Re, Ra and Ba.**

Phonetic variability for [i] vs. [a] as a function of the immediately preceding consonant is analyzed in Figure 6. For each of the pairs of sequences indicated in the figure, significant differences in degree of palatal contact and F2 frequency at the $p < 0.01$ level of significance are given. Effects are plotted at two points in time, namely, at V3 onset and V3 midpoint (EPG data), and at V3 onset/midpoint and V3 midpoint (F2 data).

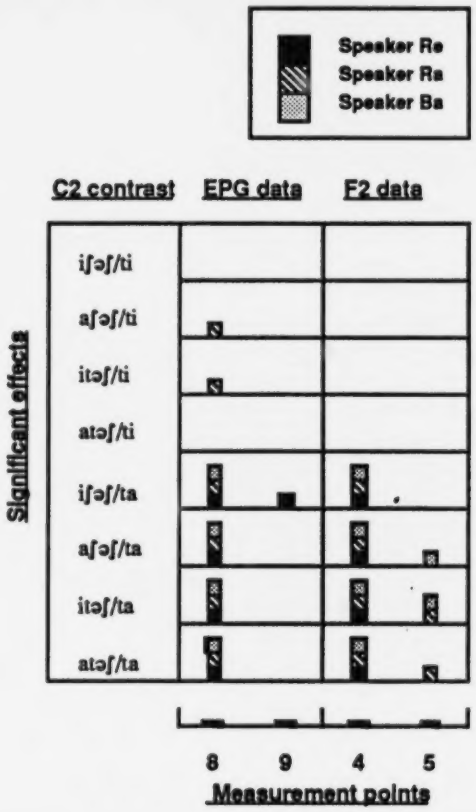


Figure 6. **EPG data and F2 data on significant coarticulatory effects upon V3=[i] and [a] from preceding C2=[ʃ] vs. [t] at the $p < 0.01$ level of significance (speakers Re, Ra and Ba). Effects are given at the V3 onset and the V3 midpoint (EPG data), and at the V3 onset/midpoint and the V3 midpoint (F2 data). See Table 2 for measurement points in time.**

The figure shows much larger consonant-dependent coarticulatory effects at the onset (EPG data) and the onset/midpoint (F2 data) of [a] than at the midpoint of [a] and along [i]. In the case of [a], these findings are due partly to the fact that the vowel onset is closer to the consonant than the vowel midpoint but also to the absence of linguopalatal contact at the midpoint of the vowel for speakers Ra and Ba. That there are no effects along V3=[i] for speakers Ra and Ba suggests that this vowel is more resistant to coarticulation than [a] because it is produced with much more palatal contact.

In summary, [i] shows a larger degree of palatal contact and is more resistant to coarticulatory effects than [a]. Data on vertical displacement for the tongue body during the production of [i] vs. [a] (Perkell & Cohen, 1987) are consistent with the findings reported here.

B. Long range coarticulatory effects

1. Vowel effects

1a. Anticipatory coarticulation

Figure 7 displays EPG and F2 data on coarticulatory effects caused by V3 along the preceding VCVC string for all speakers. The sequences in the figure were ordered for decreasing degrees of resistance to V3-dependent anticipatory effects, as explained in section II.A; thus, it was expected that these coarticulatory effects would decrease from [iʃəʃi/a] through [atəti/a] as a function of the degree of articulatory constraint associated with the VCVC string.

The EPG data show that the extent of anticipatory coarticulation is clearly dependent on the degree of palatal contact for C2 in the case of speakers Re and Ra. Overall, when C2=[ʃ], anticipatory effects may be absent or may start during C2; on the other hand, when C2=[t], V3 anticipation goes back to V2=[ə]. Therefore, two essential modes of anticipatory coarticulation appear to take place for these two speakers: later than V2=[ə] when C2=[ʃ]; at V2=[ə] when C2=[t]. These coarticulatory trends are very robust and show that the anticipatory effects are conditioned by the degree of articulatory constraint involved during the production of the preceding phonetic segment. No V3-dependent effects are available before V2=[ə]. Speaker Ba shows no effects for any of the sequences under analysis.

The F2 data are consistent to a large extent with the EPG data. For speakers Re and Ra, V3-dependent anticipatory effects take place during V2=[ə] when C2=[ʃ] and [t]; however, at least for speaker Re, C2=[ʃ] (but not C2=[t]) blocks V3-to-V2 effects in some cases. V3-dependent coarticulatory effects do not extend into V1. Speaker Ba shows almost no V3-to-V2 effects. There is no one-to-one correspondence between the EPG and the F2 data in this and in the next figures since F2 frequency is affected by changes in other articulatory dimensions besides linguopalatal contact.

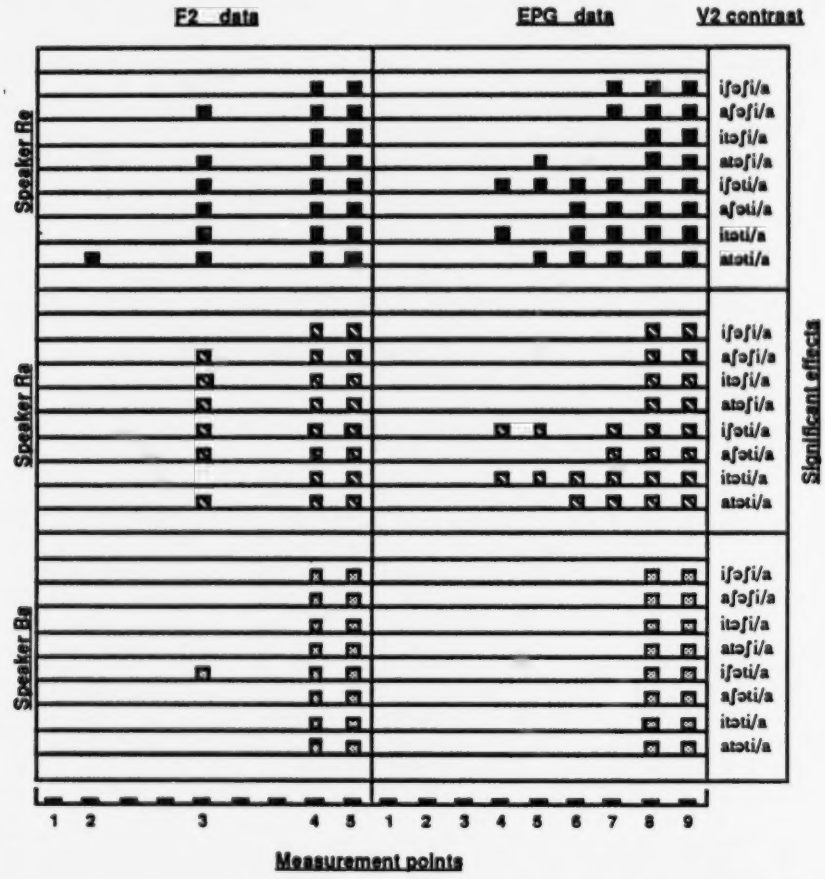


Figure 7. EPG and F2 data on significant V3-dependent anticipatory effects along the entire VC[ə]C utterance at the $p < 0.01$ level of significance (speakers Re, Ra and Ba). See Table 2 for measurement points in time.

The anticipation of the V3 gesture is not affected by the degree of coarticulatory resistance associated with the segments located in the VC string before V2=[ə]. Thus, the V3-to-V2 effects across C2=[t] take place no later when C1=[ʃ] than when C1=[t]. In fact, contrary to initial expectations, the EPG data for speakers Re and Ra show an earlier onset of anticipatory coarticulation across C2=[t] when V1=[i] than when V1=[a].

In summary, for two of the speakers, the V3-to-V2 anticipatory effects are inversely dependent on the degree of tongue dorsum contact for C2; therefore, coarticulation takes place across C2=[t] but not across C2=[ʃ]. Moreover, effects are independent of the degree of articulatory constraint for the phonetic segments preceding [ə]. No coarticulatory effects at a distance (i.e., exceeding the V3-to-V2 domain) were found; thus, the EPG and the F2 data show neither V3-to-C1 nor V3-to-V1 effects.

1b. Carryover coarticulation

Figure 8 displays EPG and F2 data on coarticulatory effects caused by V1 along the following CVCV string for all speakers. The sequences in the figure have been ordered for decreasing degrees of resistance to carryover coarticulation associated with V1, as explained in section II.A.

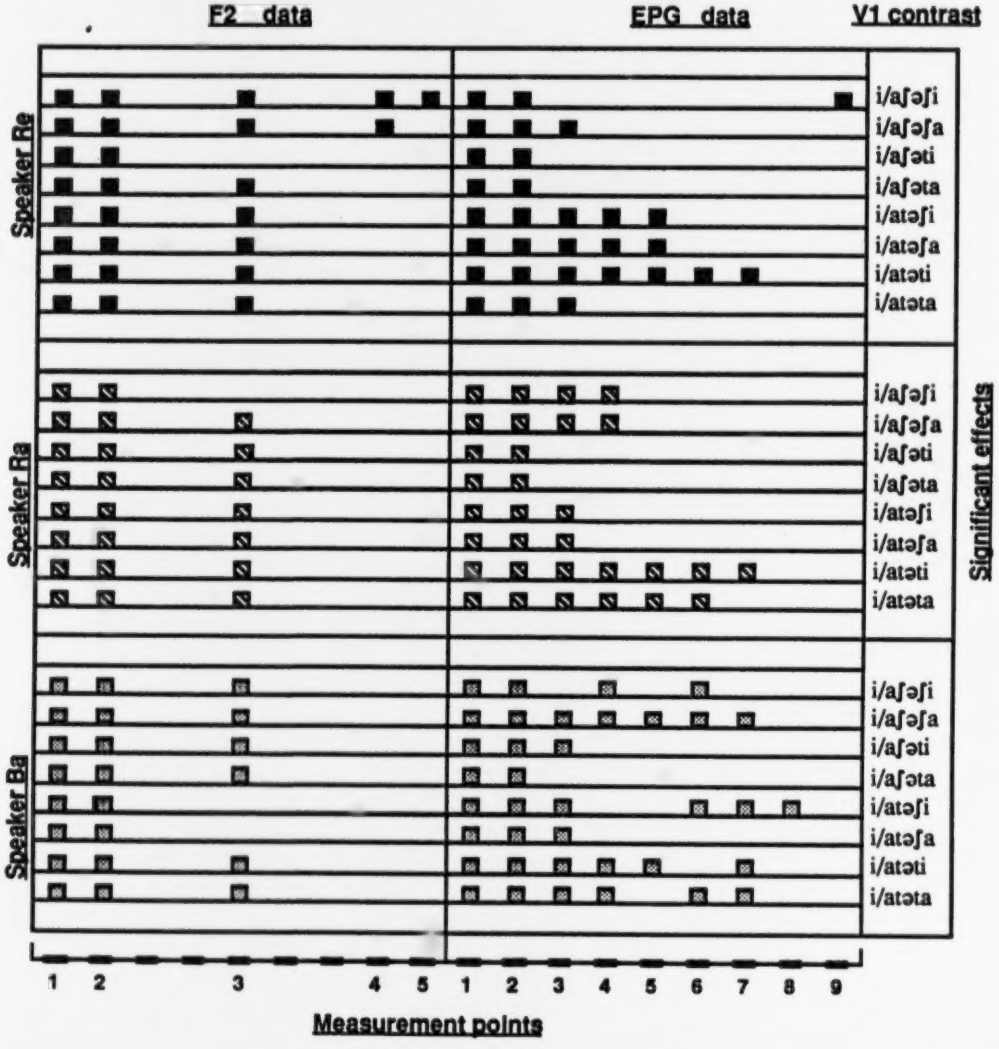


Figure 8. EPG and F2 data on significant V1-dependent carryover effects along the entire C[ə]CV utterance at the $p < 0.01$ level of significance (speakers Re, Ra and Ba). See Table 2 for measurement points in time.

The EPG data for speakers Re and Ra show clear similarities with the EPG data on anticipatory effects displayed in Figure 7. Overall, C1=[ʃ] allows lesser effects over time than C1=[t]. This pattern is not so clear for speaker Ba. Data across speakers show that the offset time of the V1-dependent carryover effects is not as constrained as the onset time of the V3-dependent anticipatory effects; while V1-dependent effects across C1=[ʃ] may extend into [ə] (and even further), those across C1=[t] may stop at C1. Also, in contrast with anticipatory effects in Figure 7, long range temporal effects are available; thus, V1-dependent carryover effects may last until C2 (mainly when C1=[t], as expected) and, occasionally, until V3.

The F2 data reveal a general trend for carryover coarticulatory effects to extend into [ə], more so when C1=[t] than when C1=[ʃ] for speakers Re and Ra but not for speaker Ba. No V1-to-V3 effects were found except in two instances for speaker Re. The fact that these two cases occur in contextual environments which were judged to be highly resistant to carryover coarticulation (i.e., sentences [i/aʃəʃi] and [i/aʃəʃa]) suggests that additional explanations are needed.

A better account of the carryover effects reported in Figure 8 results from a more detailed analysis of the articulatory constraints involved during the production of the VCVCV sequences with C1=[ʃ]. The figure shows a consistent trend for the utterances [i/aʃəti] and [i/aʃəta] to allow lesser coarticulation than the utterances [i/aʃəʃi] and [i/aʃəʃa]; as pointed out above, this is also the case for the F2 data with regard to speaker Re. The pattern of carryover effects for the utterances [i/aʃəti] and [i/aʃəta] conforms to the pattern of V3 anticipation across C2=[ʃ] (Figure 7) in that coarticulation does not reach V2=[ə]. The fact that this pattern does not apply entirely to the sequences [i/aʃəʃi] and [i/aʃəʃa], and that moreover V1-dependent carryover effects may last longer than for the sequences with C1=[t], suggests the existence of a particular production strategy.

To investigate this issue I plotted the amount of dorsopalatal contact over time for the sequences [VʃəʃV] vs. the sequences [VʃətV] (speakers Ra and Ba) in Figure 9. As shown in the figure, V1-dependent effects from [i] vs. [a] take place during V2=[ə] when C2=[ʃ] but not when C2=[t]. For the sequences [VʃətV], an active production mechanism is required for the achievement of V2=[ə]. As compared to adjacent C1=[ʃ] and C2=[t], the production of this vowel involves a noticeable decrease in degree of dorsopalatal contact; moreover, the linguopalatal target for [ə] is highly independent of differences in the quality of V1. The production of the sequences [VʃəʃV] is characterized by a different strategy. No articulatory target for the schwa is available in this case; instead, tongue dorsum contact proceeds gradually from C1 to C2 through the vowel. The degree of linguopalatal contact stays high through the entire utterance for the [iʃVʃV] sequences, and increases from V1 to C2 in the [aʃVʃV] sequences. It can be concluded that the articulatory characteristics of C2 have a clear effect on the V1-dependent coarticulatory trends in VCVCV sequences with C1=[ʃ].

The EPG data in Figure 8 reveals that the C2V3 string may affect the V1-dependent carryover effects in other instances. According to the initial hypothesis, the sequences [VtəʃV] show smaller effects (mostly until C1 or V2) than the sequences [VtətV] (mostly until V2 or C2), clearly so for speaker Ra and, less so, for speakers Re and Ba. Figure 10 illustrates this point with data on the degree of dorsopalatal contact over time for these sequences according to speaker Ra. V1-dependent effects stop at the onset of V2=[ə] when C2=[ʃ], but extend across [ə] into C2 when C2=[t]. Therefore, a highly constrained C2 (i.e., [ʃ]) blocks V1-dependent coarticulatory effects and a C2 which is unspecified for tongue dorsum contact allow these effects to take place over a long period of time.

In summary, as for anticipatory effects, carryover effects from V1 appear to be largely dependent on the degree of palatal contact for the adjacent phonetic segment (i.e., C1). Differently from anticipatory effects, the offset of carryover coarticulation is less constrained in time, may extend beyond V2=[ə], and is dependent on the articulatory characteristics of the phonemic string on the other side of [ə].

2. Consonantal effects

2a. Anticipatory coarticulation

Figure 11 displays EPG and F2 data on coarticulatory effects caused by C2 along the preceding VCV string for all speakers. The sequences have been ordered for decreasing degrees of resistance to anticipatory effects associated with C2, as explained in section II.A.

According to the figure, C2-dependent anticipatory effects over time for speaker Ra occur more frequently for sequences ending in V3=[a] than for those ending in V3=[i].

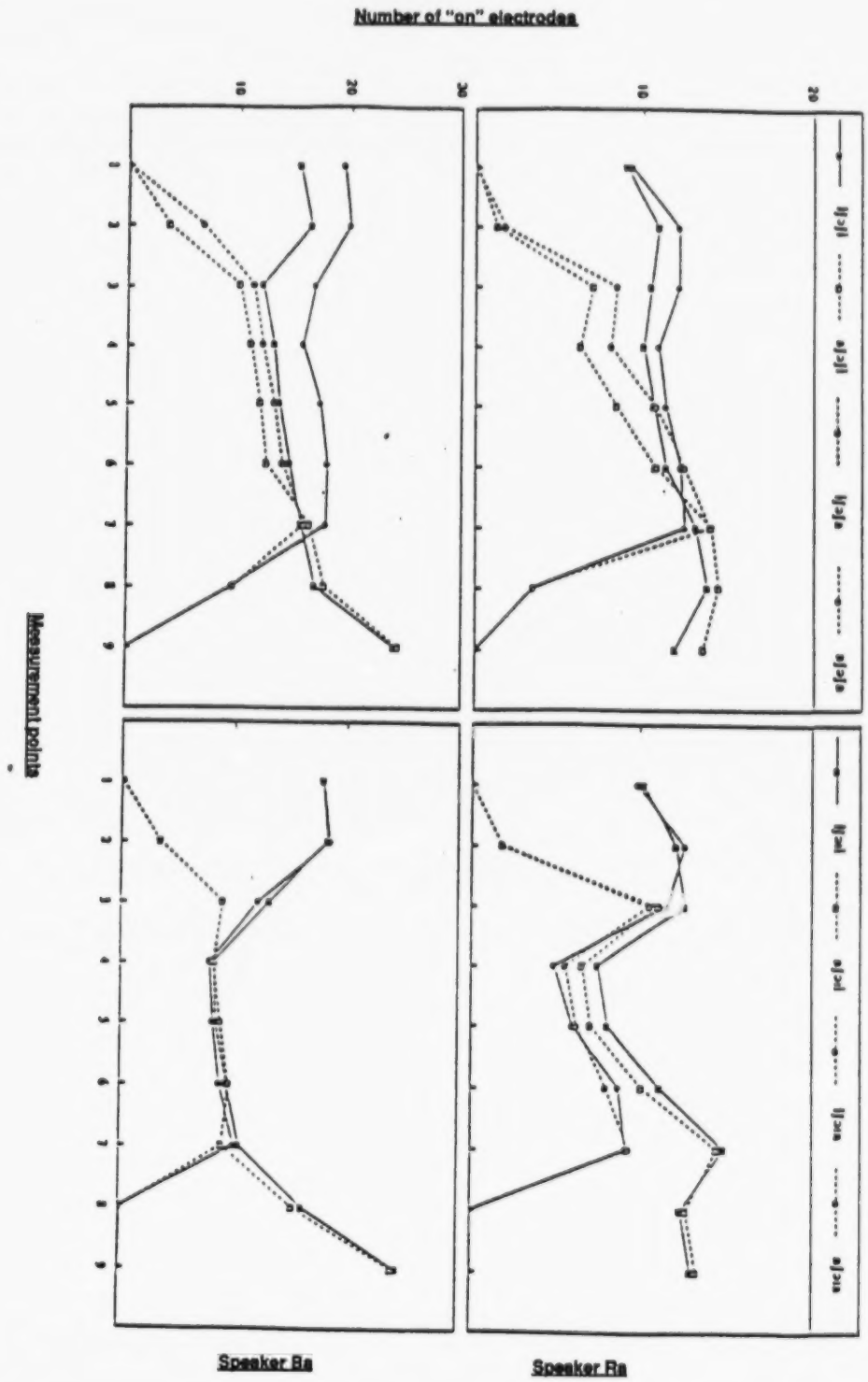


Figure 9. **V1-dependent carryover effects in tongue dorsum contact over time for the sequences [Vʃəʃ] vs. [VʃətV] according to speakers Ra and Ba. See Table 2 for measurement points in time.**

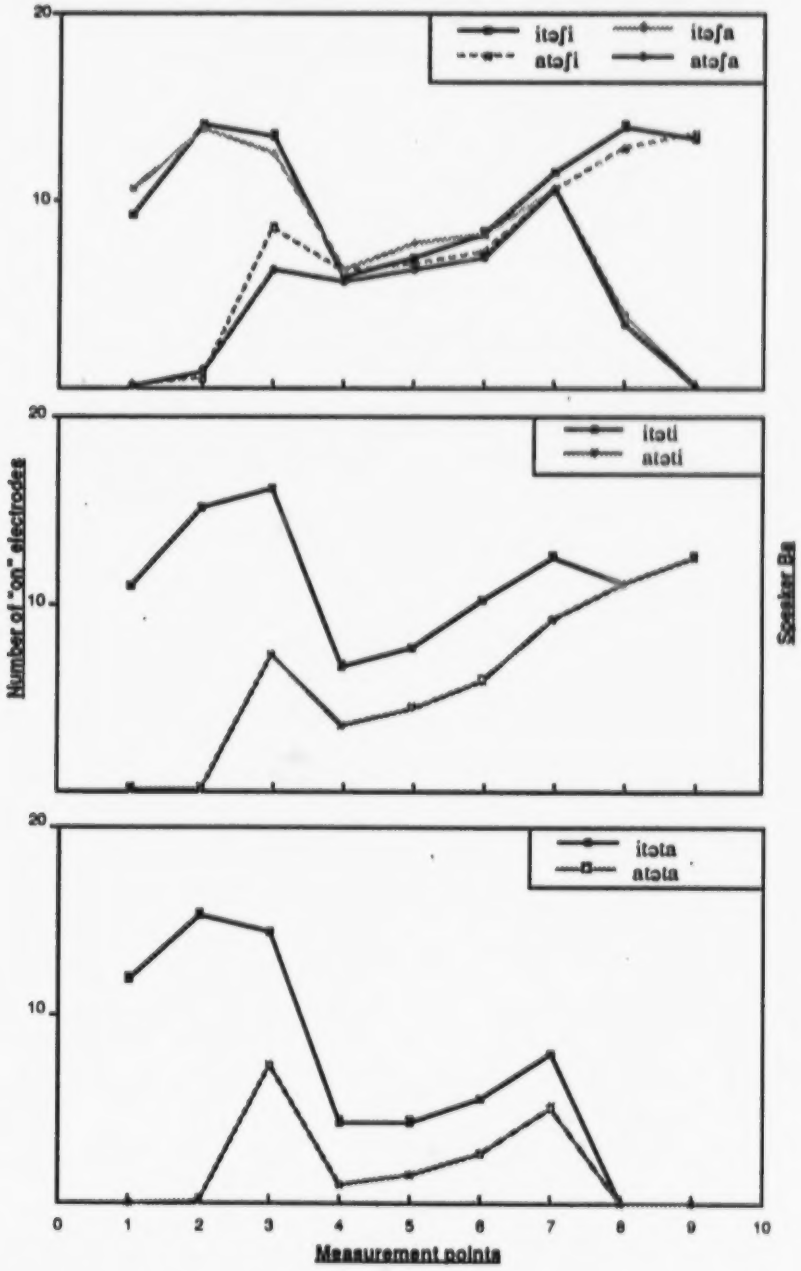


Figure 10. **V1-dependent carryover effects in tongue dorsum contact over time for the sequences [VtəʃV] vs. [VtətV] according to speaker Ra. See Table 2 for measurement points in time.**

The figure also shows that the EPG data for C2=[ʃ] and [t] before V3=[i] for the same speaker are not significantly different at the C2 midpoint (i.e., point in time 7). Anticipatory effects from C2=[ʃ] vs. [t] occur mainly when V3=[a] since there are clear differences in degree of tongue dorsum raising for the consonant in this context (see Figure 4); consonant-dependent anticipatory effects are small or non existent when V3=[i] since this vowel causes the dorsum of the tongue for [t] to show a similar amount of raising to that for [ʃ] (see Figure 4). This finding shows that the availability and extent of anticipatory coarticulation is dependent on the particular state of the articulators during the production of the target phoneme.

Overall, for all three speakers, a strong tendency is observed for C2-dependent effects to begin as early as possible during V2=[ə] but rarely during the phonetic segments preceding V2=[ə]. Effects do not appear to be dependent on the articulatory characteristics of C1; thus, for example, the onset time of the C2-dependent anticipatory effects does not occur later when C1=[ʃ] than when C1=[t]. Data for speaker Ba look somewhat different from data for speakers Re and Ra in this respect; in this case, more facilitation of the C2-dependent anticipatory effects may take place when V1=[a] than when V1=[i].

The F2 data show no C2-dependent effects before V2=[ə]. In general, those utterances showing little coarticulation in linguopalatal contact allow no C2-to-[ə] effects in F2 frequency.

In summary, as for V2-dependent anticipatory effects, C2-dependent anticipatory effects are constrained to occur at the onset of V2=[ə], more so for speakers Re and Ra than for speaker Ba. Overall, they also are highly independent of the articulatory characteristics of distant segments in the string.

2b. Carryover coarticulation

Figure 12 displays EPG and F2 data on coarticulatory effects caused by C1 along the following VCV string for all speakers. The sequences in the figure have been ordered for decreasing degrees of resistance to carryover effects associated with C1, as explained in section II.A.

Analogously to the data in Figure 11, the EPG data at the C1 midpoint show significant differences between [ʃ] and [t] after V1=[a] but not after V1=[i], for speakers Ra and Ba and, to a lesser extent, for speaker Re. As expected, the most reliable C1-dependent effects over time occur when V1=[a].

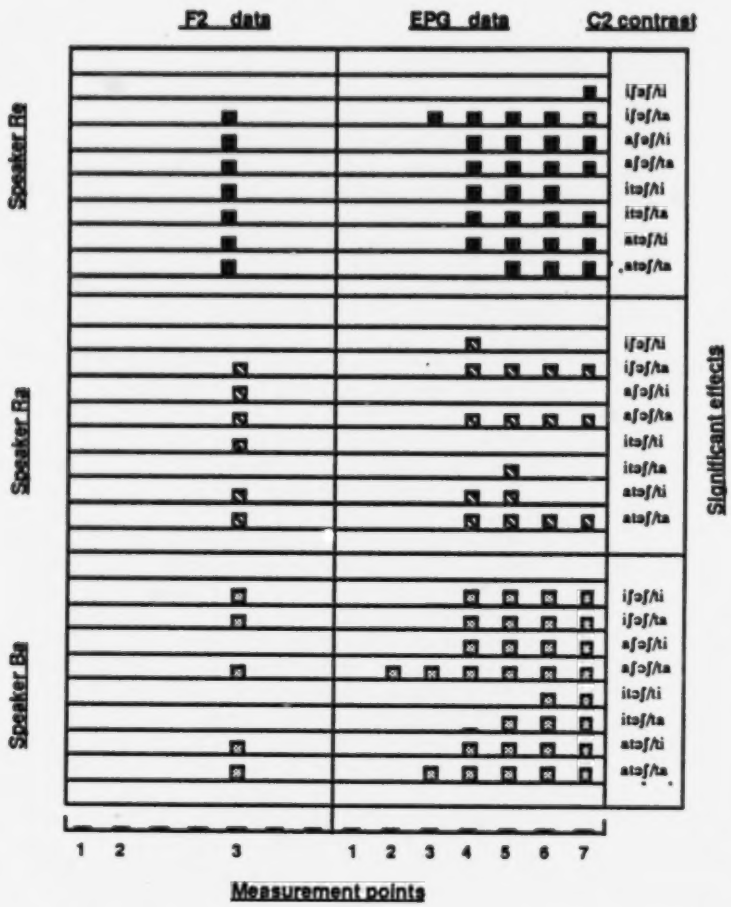


Figure 11. EPG and F2 data on significant C2-dependent anticipatory effects along the entire VC[ə] utterance at the $p < .01$ level of significance (speakers Re, Ra and Ba). See Table 2 for measurement points in time.

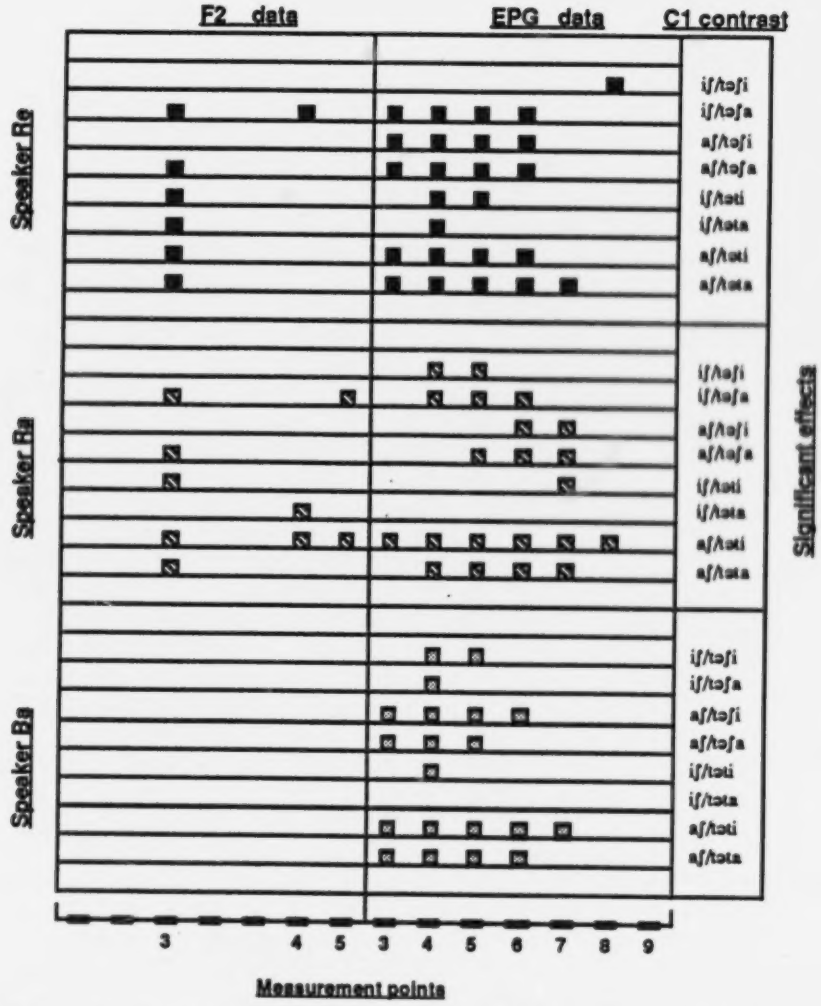


Figure 12. **EPG and F2 data on significant C1-dependent carryover effects along the entire [ə]CV utterance at the $p < .01$ level of significance (speakers Re, Ra and Ba). See Table 2 for measurement points in time.**

According to the figure, the offset time of the carryover effects appears to be affected by the articulatory characteristics of C2; thus, the EPG data (all speakers) and the F2 data (speakers Re and Ra) show a general tendency for coarticulation to last longer in the sequences [aCətV] than in the sequences [aCəʃV]. Long range effects from C1 are more salient than those from C2 in Figure 11; accordingly, the EPG and the F2 data in Figure 12 show instances of C1-to-C2 and C1-to-V3 coarticulation, mostly when C2=[t] and V3=[a], as expected.

In summary, with regard to vowel-dependent carryover vs. anticipatory effects, consonant-dependent carryover effects are more variable, may last longer and are more affected by distant phonetic segments than consonant-dependent anticipatory effects.

IV. DISCUSSION AND CONCLUSIONS

It was found in this paper that the amount of coarticulation in tongue dorsum contact varies inversely to the degree of palatal contact in adjacent or intervening segments. Data reported in the Results section (III.A.1 and III.A.2) show larger variability in tongue-dorsum contact and in F2 frequency for [t] vs. [ʃ] as a function of vocalic context, and for [i] vs. [a] as a function of consonantal context.

Data on long range coarticulatory effects show that the onset of V- and C- dependent anticipatory coarticulation is quite fixed in time, is conditioned

by the degree of tongue dorsum contact for the immediately preceding segment in the string (i.e., C2), and is speaker-dependent. Thus, the onset of the V3-dependent effects occurs during C2=[ʃ], but during preceding V2=[ə] if C=[t] (two speakers); for these same speakers the onset of the consonant-dependent effects takes place at the onset of V2=[ə]. Speaker Ba, on the other hand, shows no V3-dependent effects, and C2-dependent effects originating at the onset of [ə].

As compared with anticipatory effects, long range carryover effects from V1 and C2 were found to be quite more variable (i.e., offset times were not as regular as onset times for anticipatory effects) and to extend further in time (on C2 and V3) when distant context allowed so. Moreover, carryover effects were more sensitive than anticipatory effects to the degree of articulatory constraint for the contextual gestures. As for anticipatory coarticulation effects, they were conditioned by the degree of tongue dorsum contact for the adjacent segment (e.g., offset of vowel-dependent coarticulation occurs earlier when C1=[ʃ] than when C1=[t]). However, differently from anticipatory effects, distant segments were found to counteract or facilitate carryover coarticulation; thus, the V1-dependent effects last longer in sequences [VtətV] vs. [VtəʃV] and [VʃəʃV] vs. [VʃətV], and the offset of the C1-dependent effects takes place earlier when C2=[ʃ] than when C2=[t].

These data reveal that anticipatory and carryover effects are highly sensitive to the degree of dorsopalatal constriction exhibited by the adjacent phonetic segments in the string. Moreover, they suggest that, to a large extent, the degree of spreading of gestural activity over time may depend on some measure of degree of gestural antagonism in the adjacent and/or distant phonetic segments; it may be that the temporal domain of a gesture increases or decreases as the contextual gestures become more or less neutral, respectively, with respect to its production properties. Research reported in this study and in the recent literature (see Introduction section) provides good evidence for the notion of degree of gestural antagonism in the case of tongue dorsum raising towards the palate. Thus, the onset of gestural activity appears to take place earlier as the degree of gestural conflict for the preceding consonants decreases, for [ʃ]>[t]>[p]. Also, V1-to-V3 or V3-to-V1 effects in VCVCV sequences have been reported to occur only when V2 is highly neutral with respect to tongue body coarticulation (i.e., [ə]); consistently with this view, V-to-V effects across labial and dentoalveolar consonants in VCV sequences may not take place when the fixed vowel is not [ə] (see section I.C for references).

This issue of articulatory antagonism is crucial to elucidate the temporal domain of the coarticulatory effects (and of phonemic gestures in general), and the relative dominance of anticipatory vs. carryover coarticulation. In my view these two aspects are related as follows: the more the contextual gestures approach absolute neutrality with respect to the target gesture, the more anticipatory effects prevail over carryover effects. This situation takes place when large amounts of undershoot are available (e.g., in fast speech) and/or when preceding segments can adapt easily to the target gesture (e.g., to tongue positioning for an upcoming vowel during the production of one or more labial consonants). In such extreme circumstances anticipatory coarticulation may extend more than two segments in advance; thus, as mentioned in the Introduction section, Magen (1989) found V3-to-V1 effects in [VbəbV] sequences. Otherwise, as for /t/ in the present study (also for /l/ in Huffman, 1986) and at slow speech rates (probably in the case of speaker Ba, whose speech was the slowest of all speakers analyzed in this study), carryover effects extend over larger periods of time than anticipatory effects. Thus, in [VtətV] sequences, V- and C-dependent carryover effects were found to extend up to three (and to a lesser extent four) segments from the target and to block possible V- and C- dependent anticipatory effects before V2. I suggest that this is the case because [t] is not completely neutral with respect to tongue dorsum activity; indeed, the raising of the front of the tongue also involves some raising of the tongue dorsum.

While differences in the temporal extent of coarticulatory effects differing in directionality may be shown to depend on the degree of articulatory constraint for the adjacent gestures, a different account is needed to explain why the onset of anticipatory coarticulation is more fixed than the offset of carryover coarticulation. Thus, for example, no requirements on articulatory control may be called forth to explain why carryover effects are more dependent than anticipatory effects upon the articulatory characteristics of the distant phonemes in the string. Again, even though the articulatory configuration of V2=[ə] in the utterances of the present study is roughly equally affected by C1 and C2, the extent of V1-dependent coarticulation

is affected by the articulatory properties of C2, but the extent of V3-dependent coarticulation is not affected by the articulatory properties of C1. An obvious interpretation of this finding is that, while carryover effects may adapt to the articulatory requirements imposed upon the realization of distant phonetic segments, anticipatory effects do not do so. Instead, the onset of the anticipatory effects is required to occur at a specific moment in time or during the production of a given preceding phoneme. Together with data from other sources (Magen, 1989; Parush et al., 1983; Recasens, 1987) this may be the sort of regularity that we are looking for to state that anticipatory effects reflect motor planning to a larger extent than carryover effects.

The findings reported in this paper provide some support for a time-locked model of anticipatory coarticulation. Thus, the onset of V3-dependent effects across non-antagonistic C2=[t] was always found to occur at V2=[ə], and the onset of C2-dependent effects across non-antagonistic V2=[ə] was mostly detected at the onset of V2=[ə]. Clearly, more work is needed to determine how the onset time of articulatory activity for a given gesture changes as a function of the degree of gestural antagonism associated with the preceding phonemes in the string.

Some data reported in this and other coarticulation studies are relevant to Öhman's V-to-V model of coarticulation. Firstly, V-to-V effects barely take place if the intervening consonant is highly constrained and/or if the fixed transconsonantal vowel is not highly neutral with respect to the target vowel gesture (Recasens, 1987). Moreover, contrasting speaker-dependent behaviors may also be available; for example, it was found in the present study that V3-dependent anticipatory effects for speaker Ba did not reach the immediately preceding consonant (i.e., C2). Thirdly, as stated earlier, the temporal domain of anticipatory coarticulation may exceed two segments, if the contextual gestures are highly neutral with respect to the target gesture; moreover, the particular nature of the carryover effects also facilitates long range effects beyond V2=[ə]. Finally, as shown in figures 8 and 12 of this paper, V- and C- dependent carryover effects may land on a consonant (i.e., C2) showing a low degree of articulatory constraint. In agreement with the hypothesis of a V-to-V mode of production, in most cases V-dependent effects were found to exceed C-dependent effects. Thus, for two speakers, no C2-to-C1 anticipatory effects (even when C1=[t]) were available but only V3-to-V2 effects; of course, it can be claimed that [ə] is more neutral than [t] with regard to coarticulation in tongue dorsum activity, and that C2-to-C1 effects might have been found had C1 been [p] instead of [t]. For speaker Ba, however, while C2-to-[ə] effects occur systematically, no V3-dependent effects were found either upon V2=[ə] or even upon C2=[t]. It can be concluded that the availability of a V-to-V mode of production may vary in view of the strong and subtle dependence of coarticulation on gestural antagonism in the contextual phonemes.

REFERENCES

Alfonso P. J., & Baer, T. (1982). Dynamics of vowel articulation. *Language and Speech, 25*, 151-174.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Butcher, A., & Weiher, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics, 4*, 59-74.

Carney, P. J., & Moll, K. L. (1971). A cinefluorographic investigation of fricative consonant-vowel coarticulation. *Phonetica, 23*, 193-202.

Catford, I. (1977). *Fundamental problems in phonetics*. Bloomington: Indiana University Press.

Fant, G. (1960). *Acoustic theory of speech production*. s'Gravenhage: Mouton.

Farnetani, E., Vagges, K., & Magno-Caldognetto, E. (1985). Coarticulation in Italian /VtV/ sequences: A palatographic study. *Phonetica, 42*, 78-99.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research, 46*, 127-139.

Fowler, C. A. (1984). Current perspectives on language and speech production: A critical overview. In R. Daniloff (Ed.), *Recent advances in speech, hearing and language* (Vol. 3, pp. 195-218). San Diego: College Hill Press.

Gay, T. (1977). Cinefluorographic and electromyographic studies of articulatory organization. In M. Sawashima & F. S. Cooper (Eds.), *Dynamic aspects of speech production* (pp. 85-102). Tokyo: University of Tokyo Press.

Harris, K. S. (1984). Coarticulation as a component in articulatory description. In R. G. Daniloff (Ed.), *Articulatory assessment and treatment issues* (pp. 147-167). San Diego: College Hill Pres.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation*. Doctoral dissertation, Massachusetts Institute of Technology.

Huffman, M. K. (1986). Patterns of coarticulation in English. *UCLA Working Papers in Phonetics, 63*, 26-47.

MacNeilage, P., & DeClerk, J. L. (1969). On the motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America, 45*, 1217-1233.

McCutcheon, M., Hasegawa, A., & Fletcher, S. (1980). Effects of palatal morphology on /s, z/ articulation. *Biocommunication Research Report*, University of Alabama, Birmingham, *3*, 38-47.

Magen, H. (1989). *An acoustic study of V-to-V coarticulation in English*. Doctoral dissertation, Yale University.

Öhman, S. E. (1966). Coarticulation in VCV sequences: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*, 151-168.

Parush, A., Ostry, D. J., & Munhall, K. G. (1983). A kinematic study of lingual coarticulation in VCV sequences. *Journal of the Acoustical Society of America, 74*, 1115-1125.

Perkell, J., & Cohen, M. (1987). Token-by-token variation of tongue-body vowel targets: The effect of coarticulation. *Journal of the Acoustical Society of America, 82*, S17. (Abstract)

Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America, 76*, 1624-1635.

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in Catalan CV sequences. *Language and Speech, 28*, 97-114.

Recasens, D. (1987). An acoustical analysis of V-to-C and V-to-V coarticulatory effects in Catalan and Spanish VCV sequences. *Journal of Phonetics, 15*, 299-312.

Shibata, S., Ino, A., Yamashita, S., Hiki, S., Kiritani, S., & Sawashima, M. (1978). A new portable unit for electropalatography. *Annual Bulletin of the Institute of Logopedics and Phoniatrics*, University of Tokyo, *12*, 5-10.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular mechanics during the production of bilabial stop consonants. *Journal of Speech and Hearing Research, 16*, 397-420.

Sussman, H. M., & Westbury, J. R. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research, 46*, 16-24.

Wolf, M., Fletcher, S., McCutcheon, M., & Hasegawa, A. (1976). Medial groove width during /s/ sound production. *Biocommunication Research Reports*, University of Alabama, Birmingham, *1*, 57-66.

FOOTNOTES

*Submitted to *Speech Communication*.

†Also Universitat Autònoma de Barcelona.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 38-68

A Dynamical Approach to Gestural Patterning in Speech Production*

Elliot L. Saltzman and Kevin G. Munhall†

In this article, we attempt to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract, discrete, context-independent units, with the empirical observation of continuous, context-dependent interleaving of articulatory movements. To this end, we first review a previously proposed task-dynamic model for the coordination and control of the speech articulators. We then describe an extension of this model in which invariant speech units (gestural primitives) are identified with context-independent sets of parameters in a dynamical system having two functionally distinct but interacting levels. The *intergestural* level is defined according to a set of *activation* coordinates; the *interarticulator* level is defined according to both *model articulator* and *tract-variable* coordinates. In the framework of this extended model, coproduction effects in speech are described in terms of the blending dynamics defined among a set of temporally overlapping active units; the relative timing of speech gestures is formulated in terms of the serial dynamics that shape the temporal patterning of onsets and offsets in unit activations. Implications of this approach for certain phonological issues are discussed, and a range of relevant experimental data on speech and limb motor control is reviewed.

INTRODUCTION

The production of speech is portrayed traditionally as a combinatorial process that uses a limited set of units to produce a very large number of linguistically "well-formed" utterances (e.g., Chomsky & Halle, 1968). For example, /mæd/ and /dæm/ are characterized by different underlying sequences of the hypothesized segmental units /m/, /d/, and /æ/. These types of speech units are usually seen as discrete, static, and invariant across a variety of contexts. Putatively, such characteristics allow speech production to be generative, because units of this kind can be concatenated easily in any order to form new strings. The reality of articulation, however, bears little resemblance to this depiction. During speech production, the shape of the vocal tract changes constantly over time. These changes in shape are produced by the movements of a number of relatively independent articulators (e.g., velum, tongue, lips, jaw, etc.). For example, Figure 1 (from Krakow, 1987) displays the vertical movements of the lower lip, jaw and velum for the utterance "it's a /bamib/ sid." The lower lip and jaw cooperate to alternately close and open the mouth during /bamib/ while, simultaneously, the velum alternates between a closed and open posture. It is clear from this figure that the articulatory patterns do not take the form of discrete, abutting units that are concatenated like beads on a string. Rather, the movements of different articulators are interleaved into a continuous gestural flow. Note, for example, that velic lowering for the /m/ begins even before the lip and jaw complete the bilabial opening from the /b/ to the /a/.

We acknowledge grant support from the following sources: NIH Grant NS-13617 (Dynamics of Speech Articulation) and NSF Grant BNS-8520709 (Phonetic Structure Using Articulatory Dynamics) to Haskins Laboratories, and grants from the Natural Science and Engineering Research Council of Canada and the Ontario Ministry of Health to Kevin G. Munhall. We also thank Philip Rubin for assistance with the preparation of the Figures in this article, and Nancy O'Brien and Cathy Alfandre for their help in compiling this article's reference section. Finally, we are grateful for the critical and helpful reviews provided by Michael Jordan, Bruce Kay, Edward Reed, and Philip Rubin.

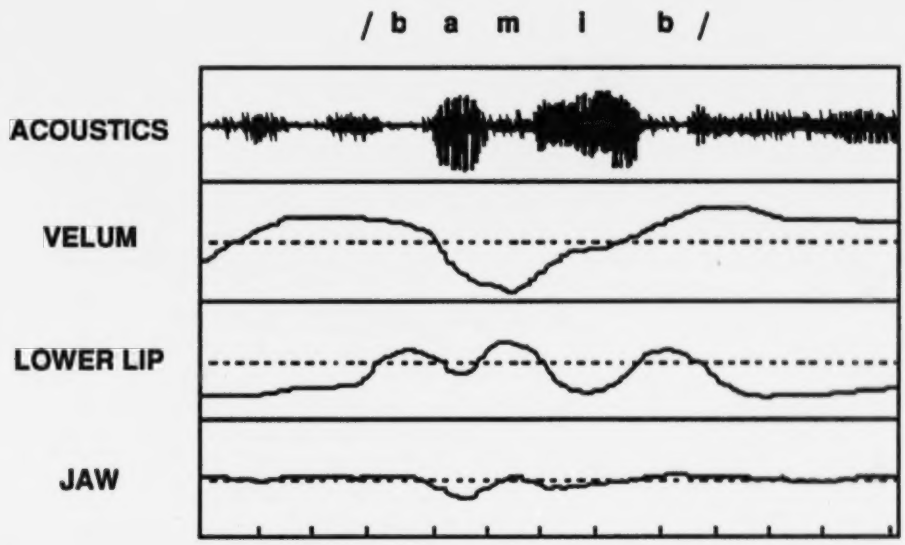


Figure 1. **Acoustic waveform and optoelectronically monitored vertical components of the articulatory trajectories accompanying the utterance "It's a /bamib/ sid". (From Krakow, 1987; used with author's permission).**

In this article, we focus on the patterning of speech gestures,[1] drawing on recent developments in experimental phonology/phonetics and in the study of coordinated behavior patterns in multi-degree-of-freedom dynamical systems. Our key questions are the following: How can one best reconcile traditional linguistic analyses (discrete, context-independent units) with experimental observations of speech articulation and acoustics (continuous, context-dependent flows)? How can one reconcile the hypothesis of underlying invariance with the reality of surface variability? We try to answer these questions by detailing a specific dynamical model of articulation. Our focus on dynamical systems derives from the fact that such systems offer a theoretically unified account of: a) the kinematic forms or patterns displayed by the articulators during speech; b) the stability of these forms to external perturbations; and c) the lawful warping of these forms due to changing system constraints such as speaking rate, casualness, segmental composition, or suprasegmental stress. For us the primary importance of the work lies not so much in the details of this model, but in the problems that can be delineated within its framework.[2] It has become clear that a complete answer to these questions will have to address (at least) the following: 1) the nature of the *gestural units* or *primitives* themselves; 2) the articulatory consequences of partial or total temporal overlap (*coproduction*) in the activities of these units that results from gestural interleaving; and 3) the *serial coupling* among gestural primitives, i.e., the processes that govern intergestural relative timing and that provide intergestural cohesion for higher-order, multigesture units.

Our central thesis is that the spatiotemporal patterns of speech emerge as behaviors implicit in a dynamical system with two functionally distinct but interacting levels. The *intergestural* level is defined according to a set of *activation* coordinates; the *interarticulator* level is defined according to both *model articulator and tract-variable* coordinates (see Figure 2). Invariant gestural units are posited in the form of relations between particular subsets of these coordinates and sets of context-independent dynamical parameters (e.g., target position and stiffness). Contextually-conditioned variability across different utterances results from the manner in which the influences of gestural units associated with these utterances are gated and blended into ongoing processes of articulatory control and coordination. The activation coordinate of each unit can be interpreted as the strength with which the associated gesture "attempts" to shape vocal tract movements at any given point in time. The tract-variable and model articulator coordinates of each unit specify the particular vocal-tract constriction (e.g., bilabial) and set of articulators (e.g., lips and jaw) whose behaviors are directly affected by the associated unit's activation. The intergestural level accounts for patterns of

relative timing and cohesion among the activation intervals of gestural units that participate in a given utterance, e.g., the activation intervals for tongue-dorsum and bilabial gestures in a vowel-bilabial-vowel sequence. The interarticulator level accounts for the coordination among articulators evident at a given point in time due to the currently active set of gestures, e.g., the coordination among lips, jaw, and tongue during periods of vocalic and bilabial gestural coproduction.[3]

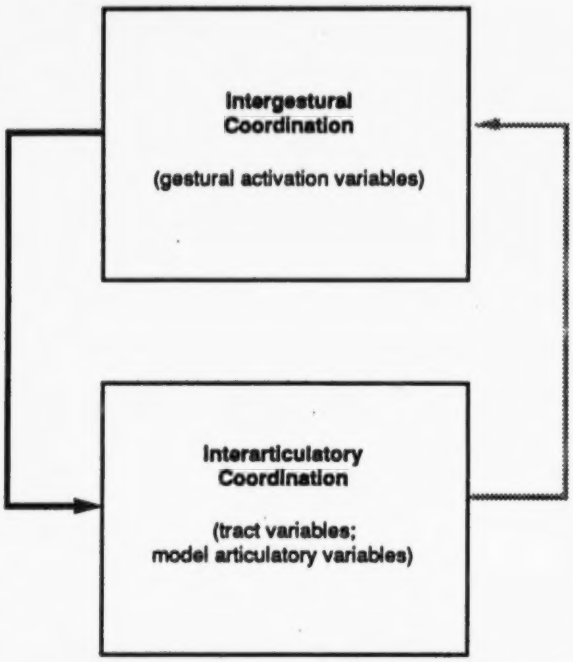


Figure 2. Schematic illustration of the proposed two-level dynamical model for speech production, with associated coordinate systems indicated. The darker arrow from the intergestural to the interarticulator level denotes the feedforward flow of gestural activation. The lighter arrow indicates feedback of ongoing tract-variable and model articulator state information to the intergestural level.

In the following pages we take a stepwise approach to elaborating upon these ideas. First, we examine the hypothesis that the formation and release of local constrictions in vocal tract shape are governed by active gestural units that serve to organize the articulators temporarily and flexibly into functional groups or ensembles of joints and muscles (i.e., synergies) that can accomplish particular gestural goals. Second, we review a recent, promising extension of this approach to the related phenomena of coarticulation and coproduction (Saltzman, Rubin, Goldstein, & Browman, 1987). Third, we describe some recent work in a connectionist, computational framework (e.g., Grossberg, 1986; Jordan, 1986, in press; Lapedes & Farber, cited in Lapedes & Farber, 1986; Rumelhart, Hinton, & Williams, 1986) that offers a dynamical account of intergestural timing. Fourth, we examine the issue of intergestural cohesion and the relationships that may exist between stable multiunit ensembles and the traditional linguistic concept of phonological segments. In doing so, we review the work of Browman and Goldstein (1986) on their *articulatory phonology*. Fifth, and finally, we review the influences of factors such as speaking rate and segmental composition on gestural patterning, and speculate on the further implications of our approach for understanding the production of speech.

Gestural primitives for speech: A dynamical framework

Much theoretical and empirical evidence from the study of skilled movements of the limbs and speech articulators supports the hypothesis that the "significant informational units of action" (Greene, 1971, p. xviii) do not entail rigid or hard-wired control of joint and/or muscle variables. Rather, these units or *coordinative structures* (e.g., Fowler, 1977; Kugler, Kelso & Turvey, 1980, 1982; Kugler & Turvey, 1987; Saltzman, 1986; Saltzman & Kelso, 1987; Turvey, 1977) must be defined abstractly or functionally in a task-specific, flexible manner. Coordinative structures have been conceptualized within the theoretical and empirical framework provided by the field of (dissipative) nonlinear dynamics (e.g., Abraham & Shaw, 1982, 1986; Guckenheimer & Holmes, 1983; Haken, 1983; Thompson & Stewart, 1986; Winfree, 1980). Specifically, it has been hypothesized (e.g., Kugler et al., 1980; Saltzman & Kelso, 1987; Turvey, 1977) that coordinative structures be defined as task-specific and autonomous (time-invariant) dynamical systems that underlie an action's form as well as its stability properties. These attributes of task-specific flexibility, functional definition, and time-invariant dynamics have been incorporated into a *task-dynamic* model of coordinative structures (Kelso, Saltzman & Tuller, 1986a, 1986b; Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman et al., 1987). In the model, time-invariant dynamical systems for specific skilled actions are defined at an abstract (task space) level of system description. These invariant dynamics underlie and give rise to contextually-

dependent patterns of change in the dynamic parameters at the articulatory level, and hence to contextually-varying patterns of articulator trajectories. Qualitative differences between the stable kinematic forms required by different tasks are captured by corresponding topological distinctions among task-space *attractors* (see also Arbib, 1984, for a related discussion of the relation between task and controller structures). As applied to limb control, for example, gestures involving a hand's discrete motion to a single spatial target and repetitive cyclic motion between two such targets are characterized by time-invariant *point attractors* (e.g., as with a damped pendulum or damped mass-spring, whose motion decays over time to a stable equilibrium point) and *periodic attractors* (*limit cycles*; e.g., as with an escapement-driven pendulum in a grandfather clock, whose motion settles over time to a stable oscillatory cycle), respectively.

Model articulator and tract variable coordinates

In speech, a major task for the articulators is to create and release constrictions locally in different regions of the vocal tract, e.g., at the lips for bilabial consonants, or between the tongue dorsum and palate for some vowels.[4] In task-dynamics, constrictions in the vocal tract are governed by a dynamical system defined at the interarticulator level (Figure 2) according to both tract variable (e.g., bilabial aperture) and model articulator (e.g., lips and jaw) coordinates. Tract variables are the coordinates in which context-*independent* gestural "intents" are framed, and model articulators are the coordinates in which context-*dependent* gestural performances are expressed. The distinction between tract-variables and model articulators reflects a behavioral distinction evident in speech production. For example, in a vowel-bilabial-vowel sequence a given degree of effective bilabial closure may be achieved with a range of different lip-jaw movements that reflects contextual differences in the identities of the flanking vowels (e.g., Sussman, MacNeilage, & Hanson, 1973).

In task-dynamic simulations, each constriction type (e.g., bilabial) is associated with a pair (typically) of tract variables, one that refers to the location of the constriction along the longitudinal axis of the vocal tract, and one that refers to the degree of constriction measured perpendicularly to the longitudinal axis in the sagittal plane. Furthermore, each gestural/constriction type is associated with a particular subset of model articulators. These simulations have been implemented using the Haskins Laboratories software articulatory synthesizer (Rubin, Baer & Mermelstein, 1981). The synthesizer is based on a midsagittal view of the vocal tract and a simplified kinematic description of the vocal tract's articulatory geometry. Modeling work has been performed in cooperation with several of our colleagues at Haskins Laboratories as part of an ongoing project focused on the development of a gesturally-based, computational model of linguistic structures (Browman & Goldstein, 1986, in press; Browman, Goldstein, Kelso, Rubin, & Saltzman, 1984; Browman, Goldstein, Saltzman, & Smith, 1986; Kelso et al., 1986a, 1986b; Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Saltzman, 1986; Saltzman et al., 1987).

Figures 3 and 4 illustrate the tract variables and articulatory degrees-of-freedom that are the focus of this article. In the present model, they are associated with the control of bilabial, tongue-dorsum, and "lower-tooth-height" constrictions.[5] Bilabial gestures are specified according to the tract variables of lip protrusion (LP; the horizontal distance of the upper and lower lips to the upper and lower front teeth, respectively) and lip aperture (LA; the vertical distance between the lips). For bilabial gestures the four modeled articulatory components are: yoked horizontal movements of the upper and lower lips (LH), jaw angle (JA), and independent vertical motions of the upper lip (ULV) and lower lip (LLV) relative to the upper and lower front teeth, respectively. Tongue-dorsum gestures are specified according to the tract variables of tongue-dorsum constriction location (TDCL) and constriction degree (TDCD). These tract variables are defined as functions of the current locations in head-centered coordinates of the region of maximum constriction between the tongue-body surface and the upper and back walls of the vocal tract. The articulator set for tongue-dorsum gestures has three components: tongue body radial (TBR) and angular (TBA) positions relative to the jaw's rotation axis, and jaw angle (JA). Lower-tooth-height gestures are specified according to a single tract variable defined by the vertical position of the lower front teeth, or equivalently, the vertical distance between the upper and lower front teeth. Its articulator set is simply jaw angle (JA). This tract variable is not used in most current simulations, but was included in the model to test hypotheses concerning suprasegmental control of the jaw (Macchi, 1985), the role of lower tooth height in tongue blade fricatives, etc.

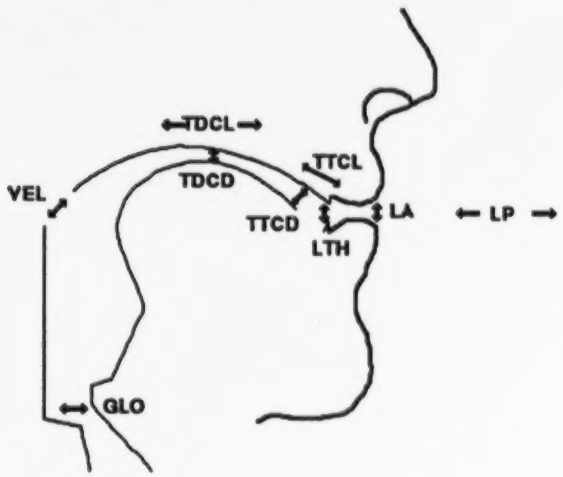


Figure 3. **Schematic midsagittal vocal tract outline, with tract-variable degrees of freedom indicated by arrows. (see text for definitions of tract-variable abbreviations used).**

MODEL ARTICULATOR VARIABLES

($\emptyset_j$; j = 1, 2, ...,n; n=10)

TRACT VARIABLES (Z_i; i = 1, 2, ...,m; m=9)	LH ($\emptyset_1$)	JA ($\emptyset_2$)	ULV ($\emptyset_3$)	LLV ($\emptyset_4$)	TBR ($\emptyset_5$)	TBA ($\emptyset_6$)	TTR ($\emptyset_7$)	TTA ($\emptyset_8$)	V ($\emptyset_9$)	G ($\emptyset_{10}$)
LP (Z_1)	●									
LA (Z_2)		●	●	●						
TDCL (Z_3)		●			●	●				
TDCD (Z_4)		●			●	●				
LTH (Z_5)		●								
TTCL (Z_6)		●			●	●	●	●		
TTCD (Z_7)		●			●	●	●	●		
VEL (Z_8)									●	
GLO (Z_9)										●

Figure 4. **Matrix representing the relationship between tract-variables (z) and model articulators ($\emptyset$). The filled cells in a given tract-variable row denote the model articulator components of that tract-variable's articulatory set. The empty cells indicate that the corresponding articulators do not contribute to the tract-variable's motion. (See text for definitions of abbreviations used in the figure.)**

Each gesture in a simulated utterance is associated with a corresponding tract-variable dynamical system. At present, all such dynamical systems are defined as tract-variable point-attractors, i.e., each is modeled currently by a damped, second-order linear differential equation (analogous to a damped mass-spring). The corresponding set of tract-variable motion equations is described in Appendix 1. These equations are used to specify a functionally equivalent dynamical system expressed in the model articulator coordinates of the Haskins articulatory synthesizer. This model articulator dynamical system is used to generate articulatory motion patterns. It is derived by transforming the tract-variable motion equations into an articulatory space whose components have geometric attributes (size, shape) but are massless. In other words, this transformation is a strictly kinematic one, and involves only the substitution of variables defined in one coordinate system for variables defined in another coordinate system (see Appendix 2).

Using the model articulator dynamical system (Equation [A4] in Appendix 2) to simulate simple utterances, the task-dynamic model has been able to generate many important aspects of natural articulation. For example, the model has been used to reproduce experimental data on *compensatory articulation*, whereby the speech system quickly and automatically reorganizes itself when faced with unexpected mechanical perturbations (e.g., Abbs & Gracco, 1983; Folkins & Abbs, 1975; Kelso, Tuller, Vatikiotis-Bateson & Fowler, 1984; Munhall & Kelso, 1985; Munhall, Löfqvist & Kelso, 1986; Shaiman & Abbs, 1987) or with static mechanical alterations of vocal tract shape (e.g., Gay, Lindblom, & Lubker, 1981; MacNeilage, 1970). Such compensation for mechanical disturbances is achieved by readjusting activity over an entire subset of articulators in a gesturally-specific manner. The task-dynamic model has been used to simulate the compensatory articulation observed during bilabial closure gestures (Saltzman, 1986; Kelso et al., 1986a, 1986b). Using point-attractor (e.g., damped mass- spring) dynamics for the control of lip aperture, when the simulated jaw is "frozen" in place during the closing gesture, at least the main qualitative features of the data are captured by the model, in that: 1) the target bilabial closure is reached (although with different final articulator configurations) for both perturbed and unperturbed "trials," and 2) compensation is immediate in the upper and lower lips to the jaw perturbation, i.e., the system does not require reparameterization in order to compensate. Significantly, in task-dynamic modeling the processes governing intra-gestural motions of a given set of articulators (e.g., the bilabial articulatory set defined by the jaw and lips) are *exactly the same* during simulations of both unperturbed and mechanically perturbed active gestures. In all cases, the articulatory movement patterns emerge as implicit consequences of the gesture-specific dynamical parameters (i.e., tract-variable parameters and articulator weights; see Appendices 1 and 2), and the ongoing postural state (perturbed or not) of the articulators. Explicit trajectory planning and/or replanning procedures are not required.

Gestural activation coordinates

Task dynamics identifies several different time spans that are important for conceptualizing the dynamics of speech production. For example, the *settling time* of an unperturbed discrete bilabial gesture is the time required for the system to move from an initial position with zero velocity to within a criterion percentage (e.g., 2%) of the distance between initial and target positions. A gesture's settling time is determined jointly by the inertia, stiffness, and damping parameters intrinsic to the associated tract-variable point attractor. Thus, gestural duration or settling time is implicit in the dynamics of the interarticulator level (Figure 2, bottom), and is not represented explicitly. There is, however, another time span that is defined by the temporal interval during which a gestural unit actively shapes movements of the articulators. In previous sections, the concept of gestural activity was used in an intuitive manner only. We now define it in a more specific fashion.

Intervals of active gestural control are specified at the intergestural level (Figure 2, top) with respect to the system's activation variables. The set of activation variables defines a third coordinate system in the present model, in addition to those defined by the tract variables and model articulators (see Figures 2 & 5). Each distinct tract-variable gesture is associated with its own activation variable, a_{ik}, where the subscript-i denotes numerically the associated tract variable ($i = 1, ..., m$), and the subscript-k denotes symbolically the particular gesture's linguistic affiliation (k = /p/, /i/, etc.). The value of a_{ik} can be interpreted as the strength with which the associated tract-variable dynamical system "attempts" to shape vocal tract movements at any given point in time.

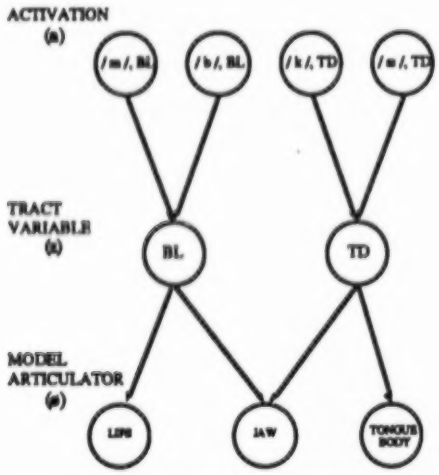


Figure 5. **Example of the "anatomical" connectivity pattern defined among the model's three coordinate systems. BL and TD denote tract-variables associated with bilabial and tongue-dorsum constrictions, respectively.**

In current simulations the temporal patterning of gestural activity is accomplished with reference to a *gestural score* (Figure 6) that represents the activation of gestural primitives over time across parallel tract-variable output channels. Currently, these activation patterns are not derived from an underlying implicit dynamics. Rather, these patterns are specified explicitly "by hand", or are derived according to a rule-based synthesis program called GEST that accepts phonetic string inputs and generates gestural score outputs (Browman et al., 1986). In the gestural score for a given utterance, the corresponding set of activation functions is specified as an explicit matrix function of time, $A(t)$. For purposes of simplicity, the activation interval of a given gesture-ik is specified according to the duration of a step-rectangular pulse in a_{ik}, normalized to unit height ($a_{ik} \in \{0, 1\}$). In future developments of the task-dynamic model (see the *Serial Dynamics* section later in the article), we plan to generalize the shapes of the activation waves and to allow activations to vary continuously over the interval ($0 \leq a_{ik} \leq 1$).[6]

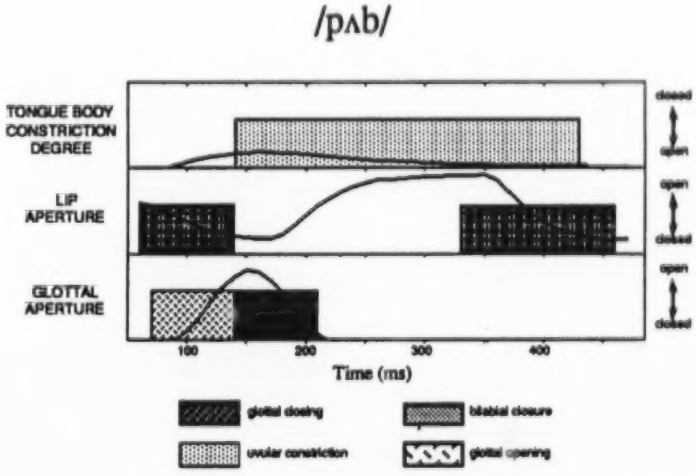


Figure 6. **Gestural score used to synthesize the sequence /pʌb/. Filled boxes denote intervals of gestural activation. Box heights are uniformly either 0 (no activation) or 1 (full activation). The waveform lines denote tract-variable trajectories generated during the simulation.**

COPRODUCTION

Temporally discrete or isolated gestures are at best rare exceptions to the rule of temporal interleaving and overlap (coproduction) among gestures associated with nearby segments (e.g., Bell-Berti & Harris, 1981; Fowler, 1977, 1980; Harris, 1984; Keating, 1985; Kent & Minifie, 1977; Öhman, 1966, 1967; Perkell, 1969; Sussman et al., 1973; for an historical review, see Hardcastle, 1981). This interleaving is the source of the ubiquitous phenomenon of coarticulation in speech production. Coarticulation refers to the fact that at any given point during an utterance, the influences of gestures associated with several adjacent or near-adjacent segments can generally be discerned in acoustic or articulatory measurements. Coarticulatory effects can occur, for example, when lip protrusion for a following rounded vowel begins during the preceding phonologically unrounded consonant, thereby coloring the acoustic correlates of the consonant with those of the following vowel. Similarly, in a vowel-/p/-vowel sequence the formation of the bilabial closure for /p/ (using the jaw and lips) appears to be influenced by temporally overlapping demands associated with the following vowel (using the jaw and tongue) by virtue of the shared jaw component.[7] In the context of the present model (see also Coker, 1976; Henke, 1966), these overlapping demands can be represented as overlapping activation patterns in a corresponding set of gestural scores. The specification of gestural scores (either by hand or by synthesis rules) thereby allows rigorous experimental control over the temporal onsets and offsets of the activations of simulated gestures, and provides a powerful computational framework and research tool for exploring and testing hypotheses derived from current work in experimental phonology/phonetics. In particular, these methods have facilitated the exploration of coarticulatory phenomena that have been ascribed to the effects of partial overlap or coproduction of speech gestures. We now describe in detail how gestural activation is incorporated into ongoing control processes in the model, and the effects of coproduction in shaping articulatory movement patterns.

Active gestural control: Tuning and gating

How might a gesture gain control of the vocal tract? In the present model, when a given gesture's activation is maximal (arbitrarily defined as 1.0), the gesture exerts maximal influence on all the articulatory components associated with the gesture's tract-variable set. During each such activation interval, the evolving configuration of the model articulators results from the gesturally- and posturally-specific way that driving influences generated in the tract-variable space (Equation [A1], Appendix 1) are distributed across the associated sets of articulatory components (Equations [A3] and [A4], Appendix 2) during the course of the movement. Conversely, when the gesture's activation is minimal (arbitrarily defined as 0.0), none of the articulators are subject to active control influences from that gesture. What, then, happens to the model articulators when there is no active control? We begin by considering the former question of active control, and treat the latter issue of nonactive control below in the section *Nonactive Gestural Control*.

The driving influences associated with a given gesture's activation "wave" (see Figure 6) are inserted into the interarticulator dynamical system in two ways in our current simulations. The first way serves to define or *tune* the current set of dynamic parameter values in the model (i.e., K, B, $\mathbf{z}_0$, and W in Equations [A3] and [A4], Appendix 2; see also Saltzman & Kelso, 1983, 1987 for a related discussion of parameter tuning in the context of skilled limb actions). The second way serves to implement or *gate* the current pattern of tract-variable driving influences into the appropriate set of articulatory components. The current use of tuning and gating is similar to Bullock and Grossberg's (1988a, 1988b; see also Cohen, Grossberg & Stork, 1988; Grossberg, 1978) use of target specification and "GO signals," respectively, in their model of sensorimotor control.

The details of tuning and gating processes depend on the ongoing patterns of overlap that exist among the gestures in a given utterance. The gestural score in Figure 6 captures in a descriptive sense both the temporal overlap of speech gestures as well as a related spatial type of overlap. As suggested in the figure, coproduction occurs whenever the activations of two or more gestures overlap partially (or wholly) in time within and/or across tract-variables. Spatial overlap occurs whenever two or more coproduced gestures share some or all of their articulatory components. In these cases, the influences of the spatially and temporally overlapping gestures are said to be *blended*. For example, in a vowel-consonant-vowel (VCV) sequence, if one assumes that the activation intervals of the vowel and

consonant gestures overlap temporally (e.g., Öhman, 1967; Sussman et al., 1973), then one can define a continuum of supralaryngeal overlap that is a function of the gestural identity of the medial consonant. In such sequences, the flanking vowel gestures are defined, by hypothesis, along the tongue-dorsum tract variables and the associated articulatory set of jaw and tongue body. If the consonant is /h/, then there is no supralaryngeal overlap. If the consonant is /b/, then its gesture is defined along the bilabial tract variables and the associated lips-jaw articulator set. Spatial overlap occurs in this case at the shared jaw. If the consonant is the alveolar /d/, its gesture is defined along the tongue-tip tract variables and the associated articulator set of tongue tip, tongue body, and jaw. Spatial overlap occurs then at the shared jaw and tongue body. Note that in both the bilabial and alveolar instances the spatial overlap is not total, and there is at least one articulator free to vary, adaptively and flexibly, in ways specific to its associated consonant. Thus, Öhman (1967) showed, for a medial alveolar in a VCV sequence, that both the location and degree of tongue-tip constriction were unaffected by the identity of the flanking vowels, although the tongue-dorsum's position was altered in a vowel-specific manner. Finally, if the medial consonant in a VCV sequence is the velar /g/, the consonant gesture is defined along exactly the same set of tract variables and articulators as the flanking vowels. In this case, there is total spatial overlap, and the system shows a loss of behavioral flexibility. That is, there is now contextual variation evident even in the attainment of the consonant's tongue-dorsum constriction target; Öhman (1967), for example, showed that in such cases the velar's place of constriction was altered by the flanking vowels, although the constriction degree was unaffected.

Blending due to spatial and temporal overlap occurs in the model as a function of the manner in which the current gestural activation matrix, A, is incorporated into the interarticulator dynamical system. Thus, blending is implemented with respect to both the gestural parameter set (tuning) and the transformation from tract-variable to articulator coordinates (gating) represented in Equations (A3) and (A4). In the following paragraphs, we describe first the computational implementation of these activation and blending processes, and then describe the results of several simulations that demonstrate their utility.

Parameter tuning. Each distinct simulated gesture is linked to a particular subset of tract-variable and articulator coordinates, and has associated with it a set of time-invariant parameters that are likewise linked to these coordinate systems. For example, a tongue-dorsum gesture's stiffness, damping, and target parameters are associated with the tract variables of tongue-dorsum constriction location and degree; its articulator weighting parameters are associated with the jaw angle, tongue-body radial, and tongue-body angular degrees of freedom. Values for these parameters are estimated from kinematic speech data obtained by optoelectronic or X-ray measurements (e.g., Kelso et al., 1985; Smith, Browman, & McGowan, 1988; Vatikiotis-Bateson, 1988). The parameter set for a given gesture is represented as:

$$k^{\dagger}_{ik},\ b^{\dagger}_{ik},\ z^{\dagger}_{oik},\ w^{\dagger}_{ikj},$$

where the subscripts denote numerically either tract variables (i = 1, ..., m) or articulators (j = 1, ..., n), or denote symbolically the particular gesture's linguistic affiliation (k = /p/, /l/, etc.). These parameter sets are incorporated into the interarticulator dynamical system (see Equations [A3] and [A4], Appendix 2) as functions of the current gestural activation matrix, A, according to explicit algebraic blending rules. These rules define or tune the current values for the corresponding components of the vector $\boldsymbol{z}_o$ and matrices K, B, and W in Equations (A3) and (A4) as follows:

$$b_{ii} = \sum_{k \in Z_i} (p_{Tik} b^{\dagger}_{ik}) \ ; \tag{1a}$$

$$k_{ii} = \sum_{k \in Z_i} (p_{Tik} k^{\dagger}_{ik}) \ ; \tag{1b}$$

$$z_{oi} = \sum_{k \in Z_i} (p_{Tik} z^{\dagger}_{oik}) \ ; \text{ and} \tag{1c}$$

$$w_{jj} = \sum_{i \in \Phi_j} \left(\sum_{k \in Z_i} p_{Wikj} w^{\dagger}_{ikj} \right) + \varepsilon N_{jj} \ , \tag{1d}$$

where p_{Tik} and p_{Wikj} are variables denoting the *post*-blending strengths of gestures whose activations influence the ongoing values of tract-variable and articulatory-weighting parameters, respectively, in Equations A1 - A4 (Appendices 1 and 2); Z_i is the set of gestures associated with the i^{th} tract-variable; Φ_j is the set of tract-variables associated with the j^{th} model articulator (see Figure 4); and $g_{Njj} = 1.0 - \min\left[1, \sum_{i\in\Phi_j}\left(\sum_{k\in Z_i} a_{ik}\right)\right]$. The subscript N denotes the fact that, for parsimony's sake, g_{Njj} are the same elements used to gate in the j^{th} articulatory component of the *neutral attractor* (see the *Nonactive Gestural Control* section to follow). In Equations (1a-1c), the parameters of the i^{th} tract-variable assume default values of zero at times when there are no active gestures that involve this tract-variable. Similarly, in Equation (1d), the articulatory weighting parameter of the j^{th} articulator assumes a default value of 1.0, due to the contribution of the g_{Njj} term, at times when there are no active gestures involving this articulator.

The p_{Tik} and p_{Wikj} terms in Equation 1 are given by the steady-state solutions of a set of feedforward, competitive-interaction-network dynamical Equations (see Appendix 3 for details). These solutions are expressed as follows:

$$p_{Tik} = a_{ik} \Big/ \left(1 + \beta_{ik} \sum_{\substack{l\in Z_i \\ l\neq k}} [\alpha_{il} a_{il}]\right); \tag{2a}$$

$$p_{Wikj} = a_{ik} \Big/ \left(1 + \beta_{ik} \sum_{i\in\Phi_j}\left[\sum_{\substack{l\in Z_i \\ l\neq k}} \alpha_{il} a_{il}\right]\right), \tag{2b}$$

where α_{il} = competitive interaction (lateral inhibition) coefficient from gesture-*il* to gesture-*ik*, for $l\neq k$; and β_{ik} = a "gatekeeper" coefficient that modulates the incoming lateral inhibition influences impinging on gesture-*ik* from gesture-*il*, for $l\neq k$; For parsimony, β_{ik} is constrained to equal $1.0/\alpha_{ik}$, for $\alpha_{ik} \neq 0.0$. If $\alpha = 0.0$, β_{ik} is set to equal 0.0 by convention. Implementing the blended parameters defined by Equations (1a-1c) into the dynamical system defined by Equations (A3) and (A4) creates an attractor layout or field of driving influences in tract-variable space that is specific to the set of currently active gestures. The blended parameters defined by Equation (1d) create a corresponding pattern of relative "receptivities" to these driving influences among the associated synergistic articulators in the coordinative structure.

Using the blending rules provided by Equations (1) and (2), different *forms* of blending can be specified, for example, among a set of temporally overlapping gestures defined within the same tract variables. The form of blending depends on the relative sizes of the context-independent (time-invariant) α and β parameters associated with each gesture. For $a_{ik} \in [0, 1]$, three possibilities are averaging, suppressing, and adding. For the set of currently active gestures along the i^{th} tract variable, if all α's are equal and greater than zero (all β's are then equal by constraint), then the $\sum_{k\in Z_i} p_{Tik}$ is normalized to equal 1.0 and the tract-variable parameters blend by simple averaging. If the α's are unequal and greater than zero, then the $\sum_{k\in Z_i} p_{Tik}$ is also normalized to equal 1.0 and the parameters blend by a weighted averaging. For example, if gesture-*ik*'s $\alpha_{ik} = 10.0$ and gesture-il's $\alpha_{il} = 0.1$, then gesture-*ik*'s parameter values dominate or "suppress" gesture-il's parameter values in the blending process when both gestures are co-active. Finally, if all α's = 0.0, then all β's = 0.0 by convention, and the parameters in Equation (1) blend by simple addition. Currently, all gestural parameters in Equation (1) are subject to the same form of competitive blending. It would be possible at some point, however, to implement different blending forms for the different parameters, e.g., adding for targets and averaging for stiffnesses, as suggested by recent data on lip protrusion (Boyce, 1988) and laryngeal abduction (Munhall & Löfqvist, 1987).

Transformation gating. The tract-variable driving influences shaped by Equations (1) and (2) remain implicit and "disconnected" from the receptive model articulators until these influences are gated explicitly into the articulatory degrees of freedom. This gating occurs with respect to the weighted Jacobian pseudoinverse (i.e., the transformation that relates tract-variable motions

to articulatory motions) and its associated orthogonal projection operator (see Appendix 2). Specifically, J^* and I_n are replaced in Equation (A4) by gated forms, J^*_G and G_P, respectively. J^*_G can be expressed as follows:

$$J^*_G = W^{-1}J_G^T(C + [I_m - G_A])^{-1}, \quad (3a)$$

where $J_G = G_A J$, and G_A is a diagonal $m \times m$ *gating* matrix for the active tract-variable gestures. Each $g_{Aii} = min\left(\sum_{k \in Z_i} a_{ik}\right)$, where the summation is defined as in Equations (1) and (2). Each g_{Aii} multiplies the i^{th} row of the Jacobian. This row relates motions of the articulators to motions defined along the i^{th} tract variable (see Equation [A2]). When $g_{Aii} = 1$ (or 0), the i^{th} tract variable is gated into (out of) J^*_G and contributes (does not contribute) to $\tilde{\varphi}_A$, the vector of active articulatory driving influences (see Equations [A3] and [A4]); $C = J_G W^{-1} J_G^T$. C embodies the kinematic interrelationships that exist among the currently active set of tract variables. Specifically, C is defined by the set of weighted, pairwise inner products of the gated Jacobian rows. A diagonal element of C, c_{ii}, is the weighted inner product (sum of squares) of the i^{th} gated Jacobian row with itself; an off-diagonal element, c_{hi} $(h \neq i)$, is the weighted inner product (sum of products) of the h^{th} and i^{th} gated Jacobian rows. A pair of gated Jacobian rows has a (generally) nonzero weighted inner product when the corresponding tract variables are active and share some or all articulators in common; the weighted inner product of two gated Jacobian rows equals zero when the corresponding tract variables are active and share no articulators; the inner product also equals zero when one or both rows correspond to a nonactive tract variable; and I_m = a $m \times m$ identity matrix.

The gated orthogonal projection operator is expressed as follows:

$$[G_P - J^*_G J], \quad (3b)$$

where G_P = a diagonal $n \times n$ gating matrix. Each element $g_{Pjj} = min\left[1, \sum_{i \in \Phi_j}\left(\sum_{k \in Z_i} a_{ik}\right)\right]$, where the summations are defined as in Equations (1) and (2).

For example, if there are no active gestures then $G_A = 0$, $G_P = 0$, and $(C + [I_m - G_A]) = I_m$ Consequently, both J^*_G and the gated orthogonal projection operator equal zero, and $\tilde{\varphi}_A = 0$ according to Equation (A4). If active gestures occur simultaneously in all tract variables, then $G_A = I_m$, $G_P = I_n$, and $J^*_G = J^*$. That is, both the gated pseudoinverse and orthogonal projection operator are "full blown" when all tract variables are active, and $\tilde{\varphi}_A$ is influenced by the attractor layouts and corresponding driving influences defined over all the tract variables. If only a few of the tract variables are active, these terms are not full blown and $\tilde{\varphi}_A$ is only subject to driving influences associated with the attractor layout in the subspace of active tract variables.

Nonactive gestural control: The neutral attractor

We return now to the question of what happens to the model articulators when there is no active control in the model. In such cases, articulator movements are shaped by a "default" *neutral attractor*. The neutral attractor is a point attractor in model articulator space, whose target configuration corresponds to schwa /ə/ in current modeling. It is possible, however, that this neutral target may be language-specific. The articulatory degrees of freedom in the neutral attractor are uncoupled dynamically, i.e., point attractor dynamics are defined independently for all articulators. At any given point in time, the neutral attractor exerts a set of driving influences on the articulators, $\tilde{\varphi}_N$, that can be expressed as follows:

$$\tilde{\varphi}_N = G_N (- B_N \dot{\varphi} - K_N[\varphi - \varphi_{N0}]), \quad (4)$$

where φ_{N0} is the neutral target configuration; B_N and K_N are $n \times n$ diagonal damping and stiffness matrices, respectively. Because their parameters never change, parameter tuning or blending is not defined for the neutral attractor. The components, k_{Njj}, of K_N are typically defined to be equal, although they may be defined asymmetrically to reflect hypothesized differences in the biomechanical time constants of the articulators (e.g., the jaw is more sluggish [has a larger time constant] than the tongue tip). The components, b_{Njj}, of B_N are defined at present relative to the corresponding K_N components to provide critical

damping for each articulator's neutral point attractor. For parsimony's sake, B_N is also used to define the orthogonal projection vector in Equation (A4); and G_N is a $n \times n$ diagonal *gating* matrix for the neutral attractor. Each element $g_{Nij} = 1.0 - \min\left[1, \sum_{i \in \Phi_j}\left(\sum_{k \in Z_i} a_{ik}\right)\right]$, where the summations are defined as in Equations (1-3). Note that $G_N = I_n - G_P$, where I_n is the n x n identity matrix and G_P is defined in Equation (3b).

The total set of driving influences ($\vec{g}_T$) on the articulators at any given point in time is the sum of an active component ($\vec{g}_A$; see Equations [A3], [A4], and [3]) and a neutral component ($\vec{g}_N$; see Equation [4]), and is defined as follows:

$$\vec{g}_T = \vec{g}_A + \vec{g}_N \qquad (5)$$

For example, consider a time when only a tongue-dorsum gesture is active. Then g_{N11} (for LH) = g_{N33} (for ULV) = g_{N44} (for LLV) = 1.0, and g_{N22} (for JA) = g_{N55} (for TBR) = g_{N66} (for TBA) = 0. Active control will exist for the gesturally involved jaw and tongue (JA, TBR, and TBA), but the noninvolved lips (LH, ULV, and LLV) will "relax" independently from their current positions toward their neutral positions according to the specified time constants. If only a bilabial gesture is active, then the complementary situation holds, with $g_{N11} = g_{N33} = g_{N44} = 0.$, and $g_{N22} = g_{N55} = g_{N66} =$ 1.0. The jaw and lips will be actively controlled, and the tongue will relax toward its neutral configuration. When both bilabial and tongue-dorsum gestures are active simultaneously, all g_{Nij} components equal zero, $\vec{g}_N = 0$, and the neutral attractor has no influence on the articulatory movement patterns. When there is no active control, all g_{Nij} components equal one, and all articulators relax toward their neutral targets.

Simulation examples

We now describe results from several simulations that demonstrate how active and neutral control influences are implemented in the model, focusing on instances of gestural coproduction.

Parameter tuning. The form of parameter blending for speech production has been hypothesized to be tract-variable-specific (Saltzman et al., 1987). As already discussed in the *Active Gestural Control* section, Öhman (1967) showed that for VCV sequences, when the medial consonant was a velar (/g/ or /k/), the surrounding vowels appeared to shift the velar's place of constriction but not its degree. These results have been simulated (qualitatively, at least) by superimposing temporally the activation intervals for the medial velar consonant and the flanking vowels. During the resultant period of consonant-vowel coproduction, an averaging blend was implemented for tuning the tract-variable parameters of tongue-dorsum-constriction-location, and a suppressing blend (velar suppresses vowel) was implemented for tongue-dorsum-constriction-degree (see Figure 7; Saltzman et al., 1987; cf., Coker, 1976, for an alternative method of generating similar results).

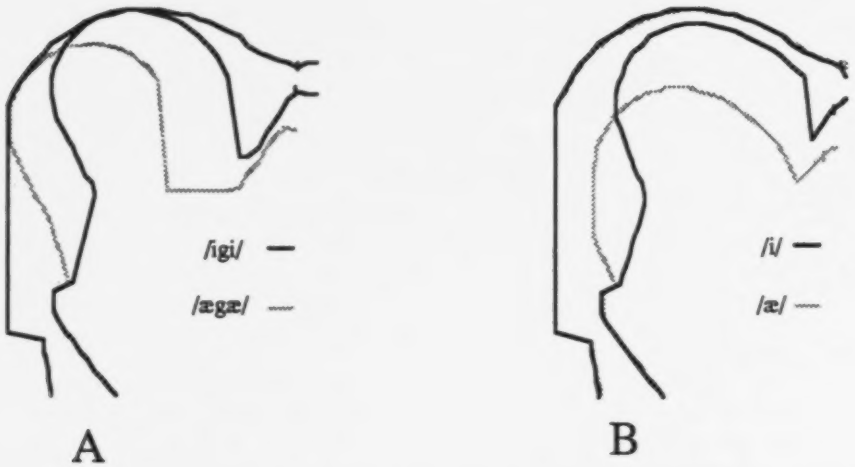


Figure 7. Simulated vocal tract shapes. A. First contact of tongue-dorsum and upper tract wall during symmetric vowel-velar-vowel sequences. B. Corresponding steady-state vowel productions.

This blending scheme for constriction degree is consistent with the assumption in current modeling that the amount of suppression during blending is related to differences in the sonority (Jesperson, 1914) or openness of the vocal tract associated with each of the blended gestures. Gestural sonority is reflected in the constriction degree target parameters of each gesture. For tongue-dorsum gestures, vowels have large positive-valued targets for constriction degree that reflect their open tract shapes (high sonority), and stops have small negative target values that reflect contact-plus-compression against the upper tract wall (low sonority).

Transformation gating. Simulations described in this article of blending for gestures defined along different tract variables have been restricted to periods of temporal overlap between pairs of bilabial and tongue-dorsum gestures. Under these circumstances, articulatory trajectories have been generated for sequences involving consonantal gestures superposed onto ongoing vocalic gestures that match (qualitatively, at least) the trajectories observed in X-ray data. In particular, Tiede and Browman (1988) analyzed X-ray data that included the vertical motions of pellets placed on the lower lip, lower incisor (i.e., jaw), and "mid-tongue" surface during /pV_1pV_2p/ sequences. The mid-tongue pellet height corresponds, roughly, to tongue-dorsum height in the current model. Tiede and Browman found that the mid-tongue pellet moved with a relatively smooth trajectory from its position at the onset of the first vowel to its position near the offset of the second vowel. Specifically, when V_1 was the medium height vowel /ɛ/ and V_2 was the low vowel /a/, the mid-tongue pellet showed a smooth lowering trajectory over this gestural time span (see Figure 8). During this same interval, the jaw and lower lip pellets moved through comparably smooth gestural sequences of lowering for V_1, raising for the medial /p/, and lowering for V_2.

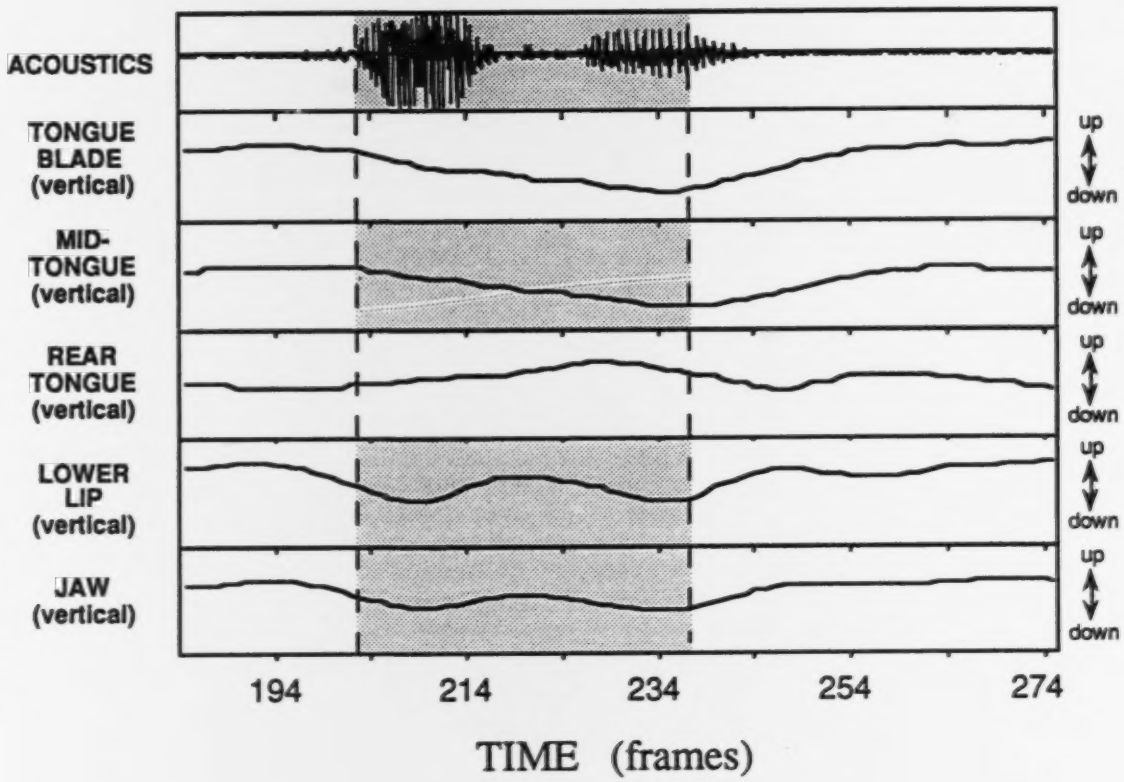


Figure 8. Acoustic waveform and vertical components of articulatory X-ray pellet data during the utterance /pɛpap/. (From Tiede & Browman, 1988; used with authors' permission).

Figure 9a shows a simulation with the current model of a similar sequence /əbæbæ/. The main point is that the vowel-to-vowel trajectory for tongue-dorsum-constriction-degree is smooth, going from the initial schwa to the more open /æ/. This tongue-dorsum pattern occurs simultaneously with the comparably smooth closing-opening gestural sequences for jaw height and lip aperture.

Two earlier versions of the present model generated nonacceptable trajectories for this same sequence that are instructive concerning the model's functioning. In one version (the "modular" model), each constriction type operated independently of the other during periods of coproduction. For example, during periods of bilabial and tongue-dorsum overlap, driving influences were generated along the tract-variables associated with each constriction. These influences were then transformed into articulatory driving influences by separate, constriction-specific Jacobian pseudoinverses (e.g., see Equations [A3] and [A4]). The bilabial pseudoinverse involved only the Jacobian rows (see Equation [A2] and Figure 4) for lip aperture and protrusion, and the tongue-dorsum pseudoinverse involved only the Jacobian rows for tongue-dorsum constriction location and degree. The articulatory driving influences associated with each constriction were simply averaged at the articulatory level for the shared jaw. The results are shown in Figure 9b, where it is evident that the tongue-dorsum does not display the relatively smooth vowel-to-vowel trajectory seen in the X-ray data and with the current model. Rather, the trajectory appears to be perturbed in a complex manner by the simultaneous jaw and lip aperture motions. It is hypothesized that these perturbations are due to the fact that the modular model did not incorporate, by definition, the off-diagonal elements of the C-matrix used currently in the gated pseudoinverse (Equation [3]). Recall that these elements reflected the kinematic relationships that exist among different, concurrently active tract-variables by virtue of shared articulators. In the modular model, these terms were absent because the constriction-specific pseudoinverses were defined explicitly to be independent of each other. Thus, if the current model is a reasonable one, it tells us that knowledge of inter-tract-variable kinematic relationships must be embodied in the control and coordinative processes for speech production.

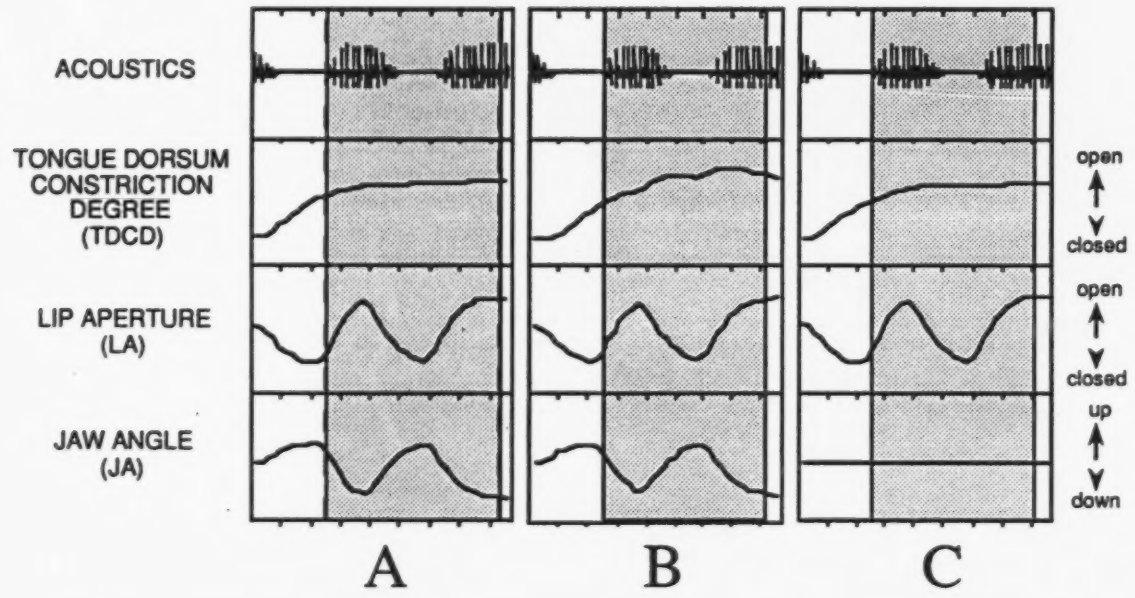


Figure 9. Simulations of the sequence /əbæbæ/. A. Current model. B. Older "modular" version. C. Older "flat jaw" version.

A different type of failure by a second earlier version of the model provides additional constraints on the form that must be taken by such inter-tract-variable knowledge. En route to developing the current model, a mistake was made that generated a perfectly flat jaw trajectory (the "flat jaw" model) for the same sequence (/əbæbæ/; see Figure 9c). Interestingly, however, the tongue-dorsum trajectory was virtually identical to that generated with the current model. The reason for this anomalous jaw behavior was that the gated pseudoinverse (Equation [3]) had been forced accidentally to be "full blown" regardless of the ongoing state of gestural activation. This meant that all tract variables were gated on in this transformation, even when the associated gestures were not activated. The specification of the attractor layout at the tract-variable level, however, worked as it does in the current model. Active gestures "create" point attractors in the control landscape for the associated tract variables. In this landscape, the currently active target can be considered to lie at the bottom of a valley whose walls are slightly sticky. The resultant tract-variable motion is to slide stably down the valley wall from its current position toward the target, due to the nonzero driving influences associated with the system's attraction to the target position. Nonactive gestures, on the other hand, "create" only flat tract-variable control landscapes, in which no position is preferred over any other and the value of the tract-variable driving influences equals zero. Recall from the *Gestural Primitives* section (Figures 3 and 4) that the model includes a lower-tooth-height tract variable that maps one-to-one onto jaw angle. For the sequence /əbæbæ/, this tract variable is never active and, consequently, the corresponding component of the tract-variable driving influence vector is constantly equal to zero. When the gated pseudoinverse is full blown, this transformation embodies the kinematic relationships among the bilabial, tongue-dorsum, and lower-tooth-height tract variables that exist by virtue of the shared jaw. This means that the transformation treats the zero driving component for lower-tooth-height as a value that should be passed on to the articulators, in conjunction with the driving influences from the bilabial and tongue-dorsum constrictions. As a result, the jaw receives zero active driving, and because the jaw starts off at its neutral position for the initial schwa, it also receives zero driving from the neutral attractor (Equation [4]) throughout the sequence. The result is the observed flat trajectory for the jaw. Thus, if the current model is a sensible one, this nonacceptable "flat jaw" simulation tells us that the kinematic interrelationships embodied in the system's pseudoinverse at any given point in time must be gated functions of the currently active gesture set.

SERIAL DYNAMICS

The task-dynamic model defines, in effect, a selective pattern of coupling among the articulators that is specific to the set of currently active gestures. This coupling pattern is shaped according to three factors: a) the current state of the gestural activation matrix; b) the tract-variable parameter sets and articulator weights associated with the currently active gestures; and c) the geometry of the nonlinear kinematic mapping between articulatory and tract-variable coordinates (represented by J and J^T in Equations [A2] and [A3]) for all associated active gestures. The model provides an intrinsically dynamical account of multiarticulator coordination within the activation intervals of single (perturbed and unperturbed) gestures. It also holds promise for understanding the blending dynamics of coproduced gestures that share articulators in common. However, task-dynamics does not currently provide an intrinsically dynamic account of the intergestural timing patterns comprising even a simple speech sequence (see Figures 1 and 6). At the level of phonologically defined segments, the sequence might be a repetitive alternation between a given vowel and consonant, e.g., /bababa.../. At a more fine-grained level of description, the sequence might be a "constellation" (Browman & Goldstein, 1986, in press) of appropriately phased gestures, e.g., the bilabial closing-opening and the laryngeal opening-closing for word-initial /p/ in English. As discussed earlier, current simulations rely on explicit gestural scores to provide the layout of activation intervals over time and tract variables for such utterances.

The lack of an appropriate *serial dynamics* is a major shortcoming in our speech modeling to date. This shortcoming is linked to the fact that the most-studied and best-understood dynamical systems in the nonlinear dynamics literature are those whose behaviors are governed by point attractors, periodic attractors (limit cycles), and *strange* attractors. (Strange attractors underlie the behaviors of *chaotic* dynamical systems, in which seemingly random movement patterns have deterministic origins; e.g., Ruelle, 1980). For nonrepetitive and nonrandom speech sequences,

such attractors appear clearly inadequate. However, investigations in the computational modeling of connectionist (parallel distributed processing, neuromorphic, neural net) dynamical systems have focused on the problem of sequence control and the understanding of serial dynamics (e.g., Grossberg, 1986; Jordan, 1986, in press; Kleinfeld & Sompolinsky, 1988; Lapedes & Farber, cited in Lapedes & Farber, 1986; Pearlmutter, 1988; Rumelhart, Hinton, & Williams, 1986; Stornetta, Hogg, & Huberman, 1988; Tank & Hopfield, 1987). Such dynamics appear well-suited to the task of sequencing or orchestrating the transitions in activation among gestural primitives in a dynamical model of speech production.

Intergestural timing: A connectionist approach

Explaining how a movement sequence is generated in a connectionist computational network becomes primarily a matter of explaining the patterning of activity over time among the network's processing elements or *nodes*. This patterning occurs through cooperative and competitive interactions among the nodes themselves. Each node can store only a small amount of information (typically only a few *marker bits* or a single scalar *activity-level*) and is capable of only a few simple arithmetic or logical actions. Consequently, the interactions are conducted, not through individual programs or symbol strings, but through very simple messages—signals limited to variations in strength. Such networks, in which the transmission of symbol strings between nodes is minimal or nonexistent, depend for their success on the availability and attunement of the right connections among the nodes (e.g., Ballard, 1986; Fahlman & Hinton, 1987; Feldman & Ballard, 1982; Grossberg, 1982; Rumelhart, Hinton, & McClelland, 1986). The knowledge constraining the performance of a serial activity, including coarticulatory patterning, is embodied in these connections rather than stored in specialized memory banks. That is, the structure and dynamics of the network govern the movement as it evolves, and knowledge of the movement's time course never appears in an explicit, declarative form.

In connectionist models, the plan for a sequence is static and timeless, and is identified with a set of input units. Output units in the network are assumed to represent the control elements of the movement components and to affect these elements in direct proportion to the level of output-unit activation. One means of producing temporal ordering is to (a) establish an activation-level gradient through lateral inhibition among the output units so that those referring to earlier aspects of the sequence are more active than those referring to later aspects; and (b) inhibit output units once a threshold value of activation is achieved (e.g., Grossberg, 1978). Such connectionist systems, however, have difficulty producing sequences in which movement components are repeated (e.g., Rumelhart & Norman, 1982). In fact, a general awkwardness in dealing with the sequential control of network activity has been acknowledged as a major shortcoming of most current connectionist models (e.g., Hopfield & Tank, 1986). Some promising developments have been reported that address such criticisms (e.g., Grossberg 1986; Jordan, 1986, in press; Kleinfeld & Sompolinsky, 1988; Lapedes & Farber, cited in Lapedes & Farber, 1986; Pearlmutter, 1988; Rumelhart, Hinton, & Willams, 1986; Stornetta, Hogg, & Huberman, 1988; Tank & Hopfield, 1987). We now describe in detail one such development (Jordan, 1986, in press).

Serial dynamics: A representative model

Jordan's (1986, in press) connectionist model of serial order can be used to define a time-invariant dynamical system with an intrinsic time scale that spans the performance of a given output sequence. There are three levels in his model (see Figure 10). At the lowest level are *output* units. Even if a particular output unit is activated repeatedly in an intended sequence, it is represented by only one unit. Thus, the model adopts a *type* rather than token representation scheme for sequence elements. In the context of the present article, a separate output unit would exist for each distinct gesture in a sequence. The tuning and gating consequences of gestural activation described earlier (see the *Active Gestural Control* section) are consistent with Jordan's suggestion that "the output of the network is best thought of as influencing articulator trajectories indirectly, by setting parameters or providing boundary conditions for lower level processes which have their own inherent dynamics" (Jordan, 1986, p. 23). For example, in the repetitive sequence /bababa.../, there would be (as a first approximation) only two output units, even though each unit potentially could be activated an indefinite number of times as the sequence continues. In this example, the

output units define the activation coordinates for the consonantal bilabial gesture and the vocalic tongue-dorsum gesture, respectively. The values of the output units are the activation values of the associated gestures, and can vary continuously across a range normalized from zero (the associated gestural unit is inactive) to one (the associated gestural unit is maximally active).

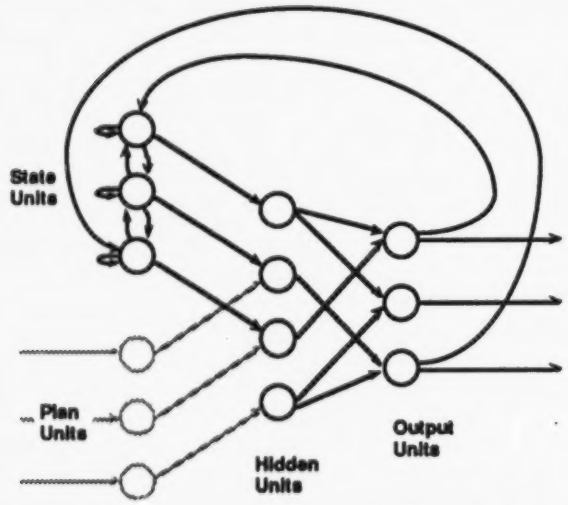


Figure 10. **Basic network architecture for Jordan's (1986, in press) connectionist model of serial order (not all connections are shown). The plan units and their connections (indicated in light grey) are not used in our proposed *hybrid* model for the serial dynamics of speech production (see text and footnote [8] for details).**

At the highest level of Jordan's model are the *state units* that, roughly speaking, define among themselves a dynamical flow with an intrinsic time scale specific to the intended sequence. These state-unit dynamics are defined by an equation of motion (the *next-state function*) that is implemented in the model by weighted recurrent connections among the state units themselves, and from the output units to the state units. Finally, at an intermediate level of the model are a set of *hidden units*. These units are connected to both the state units and the output units by two respective layers of weighted paths, thereby defining a nonlinear mapping or *output function* from state units to output units. The current vector of output activations is a function of the preceding state, which is itself a function of the previous state and previous output, and so on. Thus, the patterning over time of onsets and offsets for the output units does not arise as a consequence of direct connections among these units. Rather, such relative timing is an emergent property of the dynamics of the network as a whole. Temporal ordering among the output elements of a gestural sequence is an implicit consequence of the network architecture (i.e., the input-output functions of the system elements, and the pattern of connections among these elements) and the sequence-specific set of constant values for the weights associated with each connection path.[8]

The network can "learn" a different set of weight values for each intended utterance in Jordan's (1986; in press) model, using a "teaching" procedure that incorporates the *generalized delta rule* (*back propagation* method) of Rumelhart, Hinton, & Williams (1986). According to this rule, error signals generated at the output units (defined by the difference between the current output vector and a "teaching" vector of desired activation values) are projected back into the network to allow the hidden units to change their weights. The weights on each pathway are changed in proportion to the size of the error being back-propagated along these pathways, and error signals for each hidden unit are computed by adding the error signals arriving at these units. Rumelhart, Hinton, & Williams (1986) showed that this learning algorithm implements essentially a gradient search in weight space for the set of weights that allows the network to perform with a minimum sum of squared output errors.[9]

Jordan (1986; in press) reported simulation results in which activation values of the output units represented values of abstract phonetic features such as degree of voicing, nasality, or lip rounding. The serial network was trained to produce sequences of "phonemes", in which each phoneme was defined as a particular bundle of context-independent target values for the features. These features were not used to generate articulatory movement patterns, however. After training, the network produced continuous trajectories over time for the featural values. These trajectories displayed several impressive properties. First, the desired values were attained at the required positions in a given sequence. Second, the featural trajectories showed anticipatory and carryover coarticulatory effects for each feature that were contextually dependent on the composition of the sequence as a whole. This was due to the generalizing capacity of the network, according to which similar network states tend to produce similar outputs, and the fact that the network states during production of a given phoneme are similar to the states in which

nearby phonemes are learned. Finally, the coarticulatory temporal "spreading" of a given featural target value was not unlimited. Rather, it was restricted due to the dropoff in state similarity between a given phoneme and its surrounding context.

Toward a hybrid dynamical model

Jordan's (1986; in press) serial network has produced encouraging results for understanding the dynamics of intergestural timing in speech production. However, as already discussed, his speech simulations were defined with respect to a standard list of phonetic features, and were not related explicitly to actual articulatory movement patterns. We plan to incorporate such a serial network into our speech modeling as a means of patterning the gestural activation intervals in the task-dynamic model summarized in Equation (5). The resultant hybrid dynamical system (Figure 2) for articulatory control and coordination should provide a viable basis for further theoretical developments, guided by empirical findings in the speech production literature. For example, it is clear that the hybrid model must be able to accommodate data on the consequences for intergestural timing of mechanical perturbations delivered to the articulators during speaking. Without feedback connections that directly or indirectly link the articulators to the intergestural level, a mechanical perturbation to a limb or speech articulator could not alter the timing structure of a given movement sequence. Recent data from human subjects on unimanual oscillatory movements (Kay, 1986; Kay, Saltzman, & Kelso, 1989) and speech sequences (Gracco & Abbs, in press) demonstrate that transient mechanical perturbations induce systematic shifts in the timing of subsequent movement elements. In related animal studies (see footnote [3]), transient muscle-nerve stimulation during swimming movements of a turtle's hindlimb were also shown to induce phase shifts in the locomotor rhythm. Taken together, such data provide strong evidence that functional feedback pathways exist from the articulators to the intergestural level in the control of sequential activity. These pathways will be incorporated into our hybrid dynamical model (see the lighter pathway indicated in Figure 2).

Intrinsic vs. extrinsic timing: Autonomous vs. nonautonomous dynamics

As discussed earlier (see *Gestural Activation Coordinates* section) there are two time spans associated with every gesture in the current model. The first is the gestural settling time, defined as the time required for an idealized, temporally isolated gesture to reach a certain criterion percentage of the distance from initial to target location in tract-variable coordinates. This time span is a function of the gesture's intrinsic set of dynamic parameters (e.g., damping, stiffness). The second time-span, the gestural activation interval, is defined according to a gesture's sequence-specific activation function. In the present model, gestural activation is specified as an explicit function of time in the gestural score for a given speech sequence. In the hybrid model discussed in the previous section, these activation functions would emerge as implicit consequences of the serial dynamics intrinsic to a given sequence.

These considerations may serve to clarify certain aspects of a relatively longstanding and tenacious debate on the issue of intrinsic (e.g., Fowler, 1977, 1980) versus extrinsic (e.g., Lindblom, 1983; Lindblom et al., 1987) timing control in speech production. In the framework of the current model, intragestural temporal patterns (e.g., settling times, interarticulator asynchronies in peak velocities) can be characterized unambiguously, at least for isolated gestures, as intrinsic timing phenomena. These phenomena are emergent properties of the gesture-specific dynamics implicit in the coordinative structure spanning tract-variable and articulator coordinates (Figure 2, interarticulator level). In terms of intergestural timing, the issue is not so clear and depends on one's frame of reference. If one focuses on the interarticulatory level, then all activation inputs originate from the "outside", and activation timing must be considered extrinsic with reference to this level. Activation timing is viewed as being controlled externally according to whatever type of clock is assumed to exist or be instantiated at the system's intergestural level. However, if one considers both levels within the same frame of reference then, by definition, the timing of activation becomes intrinsic to the system as a whole. Whether or not this expansion of reference frame is useful in furthering our understanding of speech timing control depends, in part, on the nature of the clock posited at the intergestural level. This issue of clock structure leads us to a somewhat more technical consideration of the relationship between intrinsic and extrinsic timing on the one hand, and autonomous and nonautonomous dynamical systems on the other hand.

For speech production, one can posit that intrinsic timing is identified with autonomous dynamics, and extrinsic timing with nonautonomous dynamics. In an autonomous dynamical system, the terms in the corresponding equation of motion are explicit functions only of the system's "internal" state variables (i.e., positions and velocities). In contrast, a nonautonomous system's equation of motion contains terms that are explicit functions of "external" clock-time, t, such as $f(t) = \cos(\omega t)$ (e.g., Haken, 1983; Thompson & Stewart, 1986). However, the autonomous-nonautonomous distinction is just as susceptible to one's selected frame of reference as is the distinction between intrinsic and extrinsic timing. The reason is that any nonautonomous system of equations can be transformed into an autonomous one by adding an equation(s) describing the dynamics of the (formerly) external clock-time variable. That is, the frame of reference for defining the overall system equation can be extended to include the dynamics of both the original nonautonomous system as well as the formerly external clock. In this new set of equations, a state of *uni*directional coupling exists between system elements. The clock variable affects, but is unaffected by, the rest of the system variables. However, when such unidirectional coupling exists and the external clock meters out time in the standard, linear time-flow of everyday clocks and watches, we feel that its inclusion as an extra equation of motion adds little to our understanding of system behavior. In these cases, the nonautonomous description probably should be retained.

In earlier versions of the present model (Kelso et al., 1986a & 1986b; Saltzman, 1986; see also Appendices 1 & 2) only temporally isolated gestures or perfectly synchronous gesture pairs were simulated. In these cases, the equations of motion were truly autonomous, because the parameters at the interarticulatory level did not vary over the time course of the simulations. The parameters in the present model, however, are time-varying functions of the activation values specified at the intergestural level in the gestural score. Hence, the interarticulatory dynamics (Equation [5]) are currently nonautonomous. Because the gestural score specifies these activation values as explicit functions of standard clock-time, little understanding is to be gained by conceptualizing the system as an autonomous one that incorporates the unidirectionally coupled dynamics of standard clock-time and the interarticulatory level. Thus, the present model most sensibly should be considered as nonautonomous. This would not be true, however, for the proposed hybrid model in which: a) clock-time dynamics are nontrivial and intrinsic to the utterance-specific serial dynamics of the intergestural level; and b) the intergestural and interarticulator dynamics mutually affect one another. In this case, we posit that much understanding is to be gained by incorporating the dynamics of both levels into a single set of bidirectionally coupled, autonomous system equations.

INTERGESTURAL COHESION

As indicated earlier in this article (e.g., in Figure 1), speech production entails the interleaving through time of gestures defined across several different articulators and tract variables. In our current simulations, the timing of activation intervals for tract-variable gestures is controlled through the gestural score. Accordingly, gestures unfold independently over time, producing simulated speech patterns much like a player piano generates music. This rule-based description of behavior in the vocal tract makes no assumptions about coordination or functional linkages among the gestures themselves. However, we believe that such linkages exist, and that they reflect the existence of dynamical coupling within certain gestural subsets. Such coupling imbues these gestural "bundles" with a temporal cohesion that endures over relatively short (e.g., sublexical) time spans during the course of an utterance.

Support for the notion of intergestural cohesion has been provided by experiments that have focused on the structure of correlated variability evidenced between tract-variable gestures in the presence of externally delivered mechanical perturbations. Correlated variability is one of the oldest concepts in the study of natural variation, and it is displayed in a system if "when slight variations in any one part occur..., other parts become modified" (Darwin 1896, p. 128). For example, in unperturbed speech it is well known that a tight temporal relation exists between the oral and laryngeal gestures for voiceless obstruents (e.g., Löfqvist & Yoshioka, 1981a). For example, word-initial aspirated /p/ (in English) is produced with a bilabial closing-opening gesture and an accompanying glottal opening-closing gesture whose peak coincides with stop release. In a perturbation study on voiceless obstruents (Munhall et al., 1986), laryngeal compensations occurred when the lower lip was perturbed during

the production of the obstruent. Specifically, if the lower lip was unexpectedly pulled downward just prior to oral closure, the laryngeal abduction gesture for devoicing was delayed. Shaiman and Abbs (1987) have also reported data consistent with this finding. Such covariation patterns indicate a temporal cohesion among gestures, suggesting to us the existence of higher order, multigesture units in speech production.

How might intergestural cohesion be conceptualized? We hypothesize that such temporal stability can be accounted for in terms of dynamical coupling structure(s) that are defined among gestural units. Such coupling has been shown previously to induce stable intergestural phase relations in a model of two coupled gestural units whose serially repetitive (oscillatory) dynamics have been explored both experimentally and theoretically in the context of rhythmic bimanual movements (e.g., Haken, Kelso, & Bunz, 1985; Kay, Kelso, Saltzman, & Schöner, 1987; Scholz, 1986; Schöner, Haken, & Kelso, 1986). This type of model also provides an elegant account of certain changes in intergestural phase relationships that occur with increases in performance rate in the limbs and, by extension, the speech articulators. In speech, such stability and change have been examined for bilabial and laryngeal sequences consisting of either the repeated syllable /pi/ or /ip/ (Kelso, Munhall, Tuller, & Saltzman, 1985; also discussed in Kelso et al., 1986a, 1986b). When /pi/ is spoken repetitively at a self-elected "comfortable" rate, the glottal and bilabial component gestures for /p/ maintain a stable intergestural phase relationship in which peak glottal opening lags peak oral closing by an amount that results in typical (for English) syllable-initial aspiration of the /p/. For repetitive sequences of /ip/ spoken at a similarly comfortable rate, peak glottal opening occurred synchronously with peak oral closing as is typical (for English) of unaspirated (or minimally aspirated) syllable-final /p/. When /pi/ was produced repetitively at a self-paced increasing rate, intergestural phase remained relatively stable at its comfort value. However, when /ip/ was scaled similarly in rate, its phase relation was maintained at its comfort value until, at a critical speaking rate, an abrupt shift occurred to the comfort phase value and corresponding acoustic pattern for the /pi/.

In the context of the model of bimanual movement, the stable intergestural phase values at the comfort rate and the phase shift observed with rate scaling are reflections of the dynamical behavior of nonlinearly coupled, higher-order oscillatory *modes*. This use of modal dynamics parallels the identification of tract-variables with mode coordinates in the present model (see Appendix 1). Recall that the dynamics of these modal tract-variables serve to organize patterns of cooperativity among the articulators in a gesture-specific manner (see the earlier section entitled *Model Articulator and Tract Variable Coordinates*). Such interarticulator coordination is shaped according to a coupling structure among the articulators that is "provided by" the tract-variable modal dynamics. By extension, patterns of intergestural coordination are shaped according to inter-tract-variable coupling structures "provided by" a set of even higher-order multigesture modes. Because tract-variables are defined as uncoupled in the present model (Equation [A1]), it seems clear that (some sort of) inter-tract-variable coupling must be introduced to simulate the multigesture functional units evident in the production of speech.[10]

Such multigesture units could play (at least) three roles in speech production. One possibility is a hierarchical reduction of degrees of freedom in the control of the speech articulators beyond that provided by individual tract-variable dynamical systems (e.g., Bernstein, 1967). A second, related possibility is that multigesture functional units are particularly well suited to attaining articulatory goals that are relatively inaccessible to individual (or uncoupled) gestural units. For example, within single gestures the associated synergistic articulators presumably cooperate in achieving local constriction goals in tract-variable space, and individual articulatory covariation is shaped by these spatial constraints. Coordination between tract-variable gestures might serve to achieve more global aerodynamic/acoustic effects in the vocal tract. Perhaps the most familiar of such between-tract-variable effects is that of voice onset time (Lisker & Abramson, 1964), in which subtle variations in the relative timing of laryngeal and oral gestures contribute to perceived contrasts in the voicing and aspiration characteristics of stop consonants.

The third possible role for multigesture units is that of phonological primitives. For example, in Browman and Goldstein's (1986) *articulatory phonology*, the phonological primitives are gestural *constellations* that are defined as "cohesive bundles" of tract-variable gestures. Intergestural cohesion is conceived in terms of the stability of relative phasing or spatiotemporal relations among gestures within a given

constellation. In some cases, constellations correspond rather closely to traditional segmental descriptions: for example, a word-initial aspirated /p/ (in English) is represented as a bilabial closing-opening gesture and a glottal opening-closing gesture whose peak coincides with stop release; a word-initial /s/ (in English) is represented as a tongue tip raising-lowering and a glottal opening-closing gesture whose peak coincides with mid-frication. In other cases, however, it is clear that Browman and Goldstein offered a perspective that is both linguistically radical and empirically conservative. They rejected the traditional notion of segment and allowed as phonological primitives only those gestural constellations that can be observed directly from physical patterns of articulatory movements. Thus, in some instances, segmental and constellation representations diverge. For example, a word-initial /sp/ cluster (unaspirated in English) is represented as a constellation of two oral gestures (a tongue-tip and bilabial constriction-release sequence) and a *single* glottal gesture whose peak coincides with mid-frication. This representation is based on the experimental observation that for such clusters only one single-peaked glottal gesture occurs (e.g., Lisker, Abramson, Cooper, & Schvey, 1969), and thus captures the language-specific phonotactic constraint (for English) that there is no voicing contrast for stops following an initial /s/. The gestural constellation representation of /sp/ is consequently viewed as superior to a more traditional segmental approach which might predict two glottal gestures for this sequence.

Our perspective on this issue is similar to that of Browman and Goldstein in that we focus on the gestural structure of speech. Like these authors, we assume that the underlying phonological primitives are context-independent cohesive "bundles" or constellations of gestures whose cohesiveness is indexed by stable patterns of intergestural phasing. However, we adopt a position that, in comparison with theirs, is both more conservative linguistically and more radical empirically. We assume that gestures cohere in bundles corresponding, roughly, to traditional segmental descriptions, and that these segmental units maintain their integrity in fluent speech. We view many context-dependent modifications of the gestural components of these units as emergent consequences of the serial dynamics of speech production. For example, we consider the single glottal gesture accompanying English word-initial /sp/ clusters to be a within-tract-variable blend of separate glottal gestures associated with the underlying /s/ and /p/ segments (see the following section for a detailed discussion of the observable kinematic "traces" left by such underlying gestures).

INTERGESTURAL TIMING PATTERNS: EFFECTS OF SPEAKING RATE AND SEQUENCE COMPOSITION

One of the working assumptions in this article is that gestural coproduction is an integral feature of speech production, and that many factors influence the degree of gestural overlap found for a given utterance. For example, a striking phenomenon accompanying increases in speaking rate or degree of casualness is that the gestures associated with temporally adjacent segments tend to "slide" into one another with a resultant increase in temporal overlap (e.g., Browman & Goldstein, in press; Hardcastle, 1985; Machetanz, 1989; Nittrouer, Munhall, Kelso, Tuller, & Harris, 1988). Such intergestural sliding occurs both between and within tract-variables, and is influenced by the composition of the segmental sequence as well as its rate or casualness of production. We turn now to some examples of the effects on intergestural sliding and blending of changes in speaking rate and sequence composition.

Speaking rate

Hardcastle (1985) showed with electropalatographic data that the tongue gestures associated with producing the (British English) consonant sequence /kl/ tend to slide into one another and increase their temporal overlap with experimentally manipulated increases in speaking rate. Many examples of interarticulator sliding were also identified some years ago by Stetson (1951). Stetson was interested in studying the changes in articulatory timing that accompany changes in speaking rate and rhythm. Particularly interesting are his scaling trials in which utterances were spoken at increasing rates. Figure 11 is one of Stetson's figures showing the time course of lip (L), tongue (T), and air pressure (A) for productions of "sap" at different speaking rates. As can be seen, the labial gesture for /p/ and the tongue gesture for /s/ are present throughout the scaling trial but their relative timing varies with increased speaking rate. By syllable 4 the tongue gesture for /s/ and the labial gesture for /p/ from the preceding syllable completely overlap, and syllable identity is altered from then on in the trial.

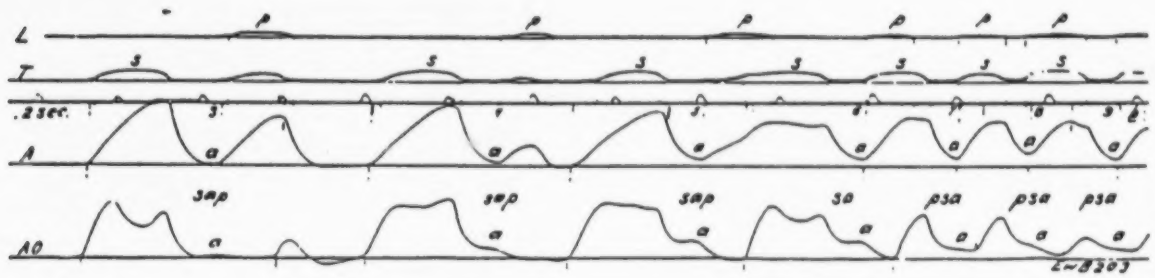


FIGURE 62. Abutting Consonants; Continuant with Stop

Syllables: *sap sap . . .*

L—Lip marker. Contact grows shorter and lighter as the rate increases and overlapping and coincidence occur.

T—Tongue marker. Well marked doubling from syl. 5-6; thereafter the single releasing compound form *ps*.

A—Air in mouth. Doubling forms, syl. 5-6.

AO—Air outside. Varied in appearance because of the high pressure during the continuant *s*. Plateau of *s* becomes mere point as compound form appears, syl. 6-7.

Figure 11. **Articulatory and aerodynamic records taken during productions of the syllable "sap" as speaking rate increases. (from Stetson, 1951; used with publisher's permission).**

In terms of the present theoretical framework, these instances of relative sliding can be described as occurring between the activation intervals associated with tongue dorsum gestures (for the velar consonant /k/), tongue tip gestures (for the alveolars /s/ and /l/), and lip aperture gestures (for the bilabial /p/). During periods of temporal overlap, the gestures sharing articulators in common are blended. Because the gestures are defined in separate tract variables, they are observably distinct in articulatory movement records. Such patterns of change might be interpretable as the response of the hybrid dynamical model discussed earlier (see *Hybrid Model* section) to hypothetically simple changes in the values of a control parameter or parameter set presumably at the model's intergestural level (see Figure 2). One goal of future empirical and simulation research is to test this notion, and if possible, to identify this parameter set and the means by which it is scaled with speaking rate.

Löfqvist & Yoshioka (1981b) have provided evidence for similar sliding and blending within tract-variables in an analysis of transillumination data on glottal devoicing gestures (abduction-adduction sequences) for a native speaker of Icelandic (a Germanic language closely related to Swedish, English, etc.). These investigators demonstrated intergestural temporal reorganization of glottal activity with spontaneous variation of speaking rate. For example, the cross-word-boundary sequence /t#k/ was accompanied by a two-peaked glottal gesture at a slow rate, but by a single-peaked gesture at fast rates. The interpretation of these data was that there were two underlying glottal gestures (one for /t/, one for /k/) at both the slow and fast rates. The visible result of only a single gesture at the fast rate appeared to be the simple consequence of blending and merging these two highly overlapping, underlying gestures defined within the same tract variable. These results have since been replicated for two speakers of North American English during experimentally controlled variations in the production rates of /s#t/ sequences (Munhall & Löfqvist, 1987).

Sequence composition: Laryngeal gestures, oral-laryngeal dominance

The rate scaling data described in the previous section for laryngeal gestures provide support for the hypothesis that the single-peaked gestures observed at fast speaking rates resulted from the sliding and blending of two underlying, sequentially adjacent gestures. In turn, this interpretation suggests a reasonable account of

glottal behavior in the production of segmental sequences containing fricative-stop clusters.

Glottal transillumination and speech acoustic data for word-final /s#ɛ/, /ks#ɛ/, and /ps#ɛ/ (unpublished data from Fowler, Munhall, Saltzman, & Hawkins, 1986a, 1986b) showed that the glottal opening-closing gesture for /s#/, in comparison to the other cases, was smaller in amplitude, shorter in duration, and peaked closer in time to the following voicing onset. These findings are consistent with the notion that a separate glottal gesture was associated with the cluster-initial stop, and that this gesture left its trace both spatially and temporally in blending with the following fricative gesture to produce a larger,[11] longer, and earlier-peaking single gestural aggregate. Other data from this experiment also indicate that the single-peaked glottal gestures observed in word-final clusters were the result of the blending of two overlapping underlying gestures. These data focus on the timing of peak glottal opening relative to the acoustic intervals (closure for /p/ or /k/, frication for /s/) associated with the production of /s#/, /ps#/, /ks#/, /sp#/, and /sk#/. For /s#/, the glottal peak occurred at mid-frication. However, for /ps#/ and /ks#/ it occurred during the first quarter of frication; for /sp#/ and /sk#/, it occurred during the third quarter of frication. These data indicate that an underlying glottal gesture was present for the /p/ or /k/ in these word-final clusters that blended with the gesture for the /s/ in a way that "pulled" or "perturbed" the peak of the gestural aggregate towards the /p/ or /k/ side of the cluster. The fact that the resultant glottal peak remained inside the frication interval for the /s/ may be ascribed, by hypothesis, to a relatively greater *dominance* over the timing of the glottal peak associated with /s/ compared to that associated with the voiceless stops /p/ or /k/.

Dominance refers to the strength of hypothesized coupling between oral acoustic/articulatory events (e.g., frication and closure intervals) and glottal events (e.g., peak glottal opening). The dominance for a voiceless consonant's oral constriction over its glottal timing appears to be influenced by (at least) two factors.[12] The first is the *manner* class of the segment: frication intervals (at least for /s/) dominate glottal behavior more strongly than stop closure intervals. This factor was highlighted previously for word-initial clusters by Browman and Goldstein (1986; cf., Kingston's [in press] related use of oral-laryngeal "binding" and Goldstein's [in press] reply to Kingston). As just discussed, this factor also appears to influence glottal timing in word-final clusters. The second factor is the presence of a preceding word-initial boundary: word-initial consonants dominate glottal behavior more strongly than the same non-word-initial consonants. These two factors appear to have approximately additive effects, as illustrated by the following examples of fricative-stop sequences defined word- or syllable-initially and across word boundaries. In these cases, as was the case word-finally, the notion of dominance can be invoked to suggest that the single-peaked glottal gestures observed for such clusters are also blends of two underlying, overlapping gestures.

Example 1. In English, Swedish, and Icelandic (e.g., Löfqvist, 1980; Löfqvist & Yoshioka, 1980, 1981a, 1981b, 1984; Yoshioka, Löfqvist, & Hirose, 1981), word-initial /s-(voiceless)stop/ clusters and /s/ are produced with a single-peaked glottal gesture that peaks at mid-frication. Word-initial /p/ is produced with a glottal gesture peaking at or slightly before closure release. Thus, the word-initial position of the /s/ in these clusters apparently bolsters the intrinsically high "segmental" dominance of the /s/, and elimi nates the displacement of the glottal peak toward the /p/ that was described earlier for word-final clusters.

Example 2. The rate scaling study for the cross-word boundary /s#t/ sequence described earlier (Munhall & Löfqvist 1987) showed two single-peaked glottal gestures for the slowest speaking rates, one double-peaked gesture for intermediate rates, and one single-peaked gesture at the fastest rate. At the slow and intermediate rates, the first peak occurred at mid-frication for the /s#/ and the second peak occurred at closure release for the /#t/. The single peak at the fastest rate occurred at the transition between frication offset and closure onset. These patterns indicate that when the two underlying glottal gestures merged into a single-peaked blend, the peak was located at a "compromise" position between the intrinsically stronger /s/ and the intrinsically weaker /t/ augmented by its word-initial status.

Example 3. In Dutch (Yoshioka, Löfqvist, & Collier, 1982), word-initial /#p/ (voiceless, unaspirated) is produced with a glottal gesture peaking at midclosure, and the glottal peak for /#s/ and /#sp/ occurs at mid-frication. However, for /#ps/ (an allowable sequence in Dutch), the glottal peak occurs at the transition between closure offset and frication onset. Again, this suggests that when the inherently stronger /s/ is augmented by word-initial status in /#sp/, the

glottal peak cannot be perturbed away from mid-frication by the following /p/. However, when the intrinsically weaker /p/ is word-initial, the glottal peak is pulled by the /p/ from mid-frication to the closure-frication boundary.

Example 4. In Swedish (e.g., Löfqvist & Yoshioka, 1980), some word-final voiceless stops are aspirated (e.g., /k#/), and are produced with glottal gestures peaking at stop release. Word-initial /#s/ is produced with glottal peak occurring at mid-frication (see Example 1). When the cross-word-boundary sequence /k#s/ is spoken at a "natural" rate, a single glottal gesture is produced with its peak occurring approximately at mid-frication. This is consistent with the high degree of glottal dominance expected for the intrinsically stronger /s/ in a word-initial position for the /k#s/ sequence.

These examples provide support for the hypothesis that fricative-stop sequences can be associated with an underlying set of two temporally overlapping but slightly offset component glottal gestures blended into a single gestural aggregate. These examples focused on the observable kinematic "traces" evident in the timing relations between the aggregate glottal peak and the acoustic intervals of the sequence. Durational data also suggest that such single observable gestures result from a two-gesture blending process. For example, the glottal gesture for the cluster /#st/ is longer in duration than either of the gestures for /#s/ and /#t/ (McGarr & Löfqvist, 1988). A similar pattern has also been found by Cooper (1989) for word-internal, syllable-initial /#s/, /#p/, and /#sp/.

SUMMARY

We have outlined an account of speech production that removes much of the apparent conflict between observations of surface variability on the one hand, and the hypothesized existence of underlying, invariant gestural units on the other hand. In doing so, we have described progress made toward a dynamical model of speech patterning that can produce fluent gestural sequences and specify articulatory trajectories in some detail. Invariant units are posited in the form of relations between context-independent sets of gestural parameters and corresponding subsets of activation, tract-variable, and articulatory coordinates in the dynamical model. Each gesture's influence over the valving and shaping of the vocal tract waxes and wanes according to the activation strengths of the units. Variability emerges in the unfolding tract-variable and articulatory movements as a result of both the utterance-specific temporal interleaving of gestural activations, and the accompanying patterns of blending or coproduction. The relative timing of the gestures and the interarticulator cooperativity evidenced for a currently active gesture set are governed by two functionally distinct but interacting levels in the model—the intergestural and interarticulatory coordination levels, respectively. At present, the dynamics of the interarticulatory level are sufficiently well developed to offer promising accounts of movement patterns observed during unperturbed and mechanically perturbed speech sequences, and during periods of coproduction. We have only begun to explore the dynamics of the intergestural level. Yet even these preliminary considerations, grounded in developments in the dynamical systems literature, have already begun to shed light on several longstanding issues in speech science, namely, the issues of intrinsic versus extrinsic timing, the nature of intergestural cohesion, and the hypothesized existence of segmental units in the production of speech. We find these results encouraging, and look forward to further progress within this research framework.

REFERENCES

Abbs, J. H., & Gracco, V. L. (1983). Sensorimotor actions in the control of multimovement speech gestures. *Trends in Neuroscience, 6*, 393-395.

Abraham, R., & Shaw, C. (1982). *Dynamics-The geometry of behavior. Part 1: Periodic behavior*. Santa Cruz, CA: Aerial Press.

Abraham, R., & Shaw, C. (1986). Dynamics: A visual introduction. In F. E. Yates (Ed.), *Self-organizing systems: The emergence of order*. New York: Plenum Press.

Arbib, M. A. (1984). From synergies and embryos to motor schemas. In H. T. A. Whiting (Ed.), *Human motor actions: Bernstein reassessed* (pp. 545-562). New York: North-Holland.

Ballard, D. H. (1986). Cortical connections and parallel processing: Structure and function. *The Behavioral and Brain Sciences, 9*, 67-120.

Ballieul, J., Hollerbach, J., & Brocket, R. (1984; December). Programming and control of kinematically redundant manipulators. *Proceedings of the 23rd IEEE Conference on Decision and Control*. Las Vegas, NV.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Benati, M., Gaglio, S., Morasso, P., Tagliasco, V., & Zaccaria, R. (1980). Anthropomorphic robotics. I. Representing mechanical complexity. *Biological Cybernetics, 38*, 125-140.

Bernstein, N. A. (1967). *The coordination and regulation of movements*. London: Pergamon Press. Reprinted in H. T. A. Whiting (Ed.). (1984). *Human motor actions: Bernstein reassessed*. New York: North-Holland.

Boyce, S. E. (1988). *The influence of phonological structure on articulatory organization in Turkish and English: Vowel harmony and coarticulation*. Unpublished doctoral dissertation, Department of Linguistics, Yale University.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (in press). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology: I. Between the Grammar and the Physics of Speech*. Cambridge, England: Cambridge University Press.

Browman, C. P., Goldstein, L., Kelso, J. A. S., Rubin, P., & Saltzman, E. L. (1984). Articulatory synthesis from underlying dynamics [Abstract]. *Journal of the Acoustical Society of America, 75* (Suppl. 1), S22-S23.

Browman, C. P., Goldstein, L., Saltzman, E. L., & Smith, C. (1986). GEST: A computational model for speech production using dynamically defined articulatory gestures [Abstract]. *Journal of the Acoustical Society of America, 80* (Suppl. 1), S97.

Bullock, D., & Grossberg, S. (1988a). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review, 95*, 49-90.

Bullock, D., & Grossberg, S. (1988b). The VITE model: A neural command circuit for generating arm and articulator trajectories. In J. A. S. Kelso, A. J. Mandell, & M. F. Schlesinger (Eds.), *Dynamic patterns in complex systems*. Singapore: World Scientific Publishers.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Cohen, M. A., Grossberg, S., & Stork, D. G. (1988). Speech perception and production by a self-organizing neural network. In Y. C. Lee (Ed.), *Evolution, learning, cognition, and advanced architectures*. Hong Kong: World Scientific Publishers.

Coker, C. H. (1976). A model of articulatory dynamics and control. *Proceedings of the IEEE, 64*, 452-460.

Cooper, A. (1989). [An articulatory description of the distribution of aspiration in English]. Unpublished research.

Darwin, C. (1896). *The origin of species*. New York: Caldwell.

Fahlman, S. E., & Hinton, G. E. (1987). Connectionist architectures for artificial intelligence. *Computer, 20*, 100-109.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 9*, 205-254.

Folkins, J. W., & Abbs, J. H. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research, 18*, 207-220.

Fowler, C. A. (1977). *Timing control in speech production*. Bloomington, IN: Indiana University Linguistics Club.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing control. *Journal of Phonetics, 8*, 113-133.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: Human Perception and Performance, 112*, 386-412.

Fowler, C. A., Munhall, K. G., Saltzman, E. L., & Hawkins, S. (1986a). Acoustic and articulatory evidence for consonant-vowel interactions [Abstract]. *Journal of the Acoustical Society of America, 80* (Suppl. 1), S96.

Fowler, C. A., Munhall, K. G., Saltzman, E. L., & Hawkins, S. (1986b). [Laryngeal movements in word-final single consonants and consonant clusters]. Unpublished research.

Gay, T. J., Lindblom, B., & Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America, 69*, 802-810.

Goldstein, L. (in press). On articulatory binding: Comments on Kingston's paper. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology: I. Between the grammar and the physics of speech*. Cambridge, England: Cambridge University Press.

Gracco, V. L., & Abbs, J. H. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research, 65*, 156-166.

Gracco, V. L., & Abbs, J. H. (1989). Sensorimotor characteristics of speech motor sequences. *Experimental Brain Research, 75*, 586-598.

Greene, P. H. (1971). Introduction. In I. M. Gelfand, V. S. Gurfinkel, S. V. Fomin, & M. L. Tsetlin (Eds.), *Models of the structural-functional organization of certain biological systems* (pp. xi-xxxi). Cambridge, MA: MIT Press.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (Vol. 5, pp. 233-374). New York: Academic Press.

Grossberg, S. (1982). *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control*. Amsterdam: Reidel Press.

Grossberg, S. (1986). The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines* (Vol. 1, pp. 187-294). New York: Academic Press.

Grossberg, S., & Mingolla, E. (1986). Computer simulation of neural networks for perceptual psychology. *Behavior Research Methods, Instruments, & Computers, 18*, 601-607.

Guckenheimer, J., & Holmes, P. (1983). *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. New York: Springer-Verlag.

Haken, H. (1983). *Advanced synergetics*. Heidelberg: Springer-Verlag.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51*, 347-356.

Hardcastle, W. J. (1981). Experimental studies in lingual coarticulation. In R. E. Asher & E. J. A. Henderson (Eds.), *Towards a history of phonetics* (pp. 50-66). Edinburgh, Scotland: Edinburgh University Press.

Hardcastle, W. J. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication, 4*, 247-263.

Harris, K. S. (1984). Coarticulation as a component of articulatory descriptions. In R. G. Daniloff (Ed.), *Articulation assessment and treatment issues* (pp. 147-167). San Diego: College Hill Press.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science*, 233, 625-633.

Jespersen, O. (1914). *Lehrbuch der Phonetik* [Textbook of Phonetics]. Leipzig: Teubner.

Joos, M. (1948). Acoustic phonetics. *Language, 24 (SM23)*, 1-136.

Jordan, M. I. (1985). *The learning of representations for sequential performance*. Unpublished doctoral dissertation, Department of Cognitive Science and Psychology, University of California, San Diego, CA.

Jordan, M. I. (1986). *Serial order in behavior: A parallel distributed processing approach* (Tech. Rep. No. 8604). San Diego: University of California, Institute for Cognitive Science.

Jordan, M.I. (in press). Serial order: A parallel distributed processing approach. In J.L. Elman & D.E. Rumelhart (Eds.), *Advances in connectionist theory: Speech*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kay, B. A. (1986). *Dynamic modeling of rhythmic limb movements: Converging on a description of the component oscillators*. Unpublished doctoral dissertation, Department of Psychology, University of Connecticut, Storrs, CT.

Kay, B. A., Kelso, J. A. S., Saltzman, E. L., & Schöner, G. (1987). Space-time behavior of single and bimanual rhythmical movements: Data and limit cycle model. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 178-192.

Kay, B. A., Saltzman, E. L., & Kelso, J. A. S. (1989). *Steady-state and perturbed rhythmical movements: A dynamical analysis*. Manuscript submitted for publication.

Keating, P. A. (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA Working Papers in Phonetics, 62*, 1-13.

Kelso, J. A. S., Munhall, K. G., Tuller, B., & Saltzman, E. L. (1985). [Phase transitions in speech production]. Unpublished research.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986a). The dynamical theory on speech production: Data and theory. *Journal of Phonetics, 14*, 29-60.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986b). Intentional contents, communicative context, and task dynamics: A reply to the commentators. *Journal of Phonetics, 14*, 171-196.

Kelso, J.A.S. & Tuller, B. (in press). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and performance, XIII*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 812-832.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay, B. A. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77*, 266-280.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5*, 115-133.

Kingston, J. (in press). Articulatory binding. In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology: I. Between the grammar and physics of speech*. Cambridge, England: Cambridge University Press.

Klein, C. A., & Huang, C. H. (1983). Review of pseudoinverse control for use with kinematically redundant manipulators. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 245-250.

Kleinfeld, D. & Sompolinsky, H. (1988). Associative neural network model for the generation of temporal patterns: Theory and application to central pattern generators. *Biophysical Journal, 54, 1039-1051*.

Krakow, R. A. (1987; November). *Stress effects on the articulation and coarticulation of labial and velic gestures*. Paper presented at the meeting of the American Speech-Language-Hearing Association, New Orleans, LA.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1980). On the concept of coordinative structures as dissipative structures: I. Theoretical lines of convergence. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 3-47). New York: North-Holland.

Kugler, P. N., Kelso, J. A. S., & Turvey, M. T. (1982). On the control and coordination of naturally developing systems. In J. A. S. Kelso & J. E. Clark (Eds.), *The development of movement control and coordination* (pp. 5-78). Chichester, England: John Wiley.

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lapedes, A., & Farber, R. (1986). *Programming a massively parallel, computation universal system: Static behavior* (Preprint No. LA-UR 86-7179). Los Alamos, NM: Los Alamos National Laboratory, .

Lennard, P. R. (1985). Afferent perturbations during "monopodal" swimming movements in the turtle: Phase-dependent cutaneous modulation and proprioceptive resetting of the locomotor rhythm. *The Journal of Neuroscience, 5*, 1434-1445.

Lennard, P. R., & Hermanson, J. W. (1985). Central reflex modulation during locomotion. *Trends in Neuroscience, 8*, 483-486.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 217-245). New York: Springer-Verlag.

Lindblom, B., Lubker, J., Gay, T., Lyberg, B., Branderud, P., & Holmgren, K. (1987). The concept of target and speech timing. In R. Channon & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 161-181). Dordrecht, Netherlands: Foris Publications.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384-422.

Lisker, L., Abramson, A. S., Cooper, F. S., & Schvey, M. H. (1969). Transillumination of the larynx in running speech. *Journal of the Acoustical Society of America, 45*, 1544-1546.

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics, 8*, 475-490.

Löfqvist, A., & Yoshioka, H. (1980). Laryngeal activity in Swedish obstruent clusters. *Journal of the Acoustical Society of America, 68*, 792-801.

Löfqvist, A., & Yoshioka, H. (1981a). Interarticulator programming in obstruent production. *Phonetica, 38*, 21-34.

Löfqvist, A., & Yoshioka, H. (1981b). Laryngeal activity in Icelandic obstruent production. *Nordic Journal of Linguistics, 4*, 1-18.

Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication, 3*, 279-289.

Macchi, M. (1985). *Segmental and suprasegmental features and lip and jaw articulators*. Unpublished doctoral dissertation, Department of Linguistics, New York University, New York, NY.

Machetanz, J. (1989, May). *Tongue movements in speech at different rates*. Paper presented at the Salk Laboratory for Language and Cognitive Studies, San Diego, CA.

MacNeilage, P. F. (1970). Motor control of serial ordering of speech. *Psychological Review, 77*, 182-196.

Mattingly, I. G. (1981). Phonetic representation and speech synthesis by rule. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 415-420). Amsterdam: North-Holland.

McGarr, N. S., & Löfqvist, A. (1988). Laryngeal kinematics in voiceless obstruents produced by hearing-impaired speakers. *Journal of Speech and Hearing Research, 31*, 234-239.

Miyata, Y. (1987). Organization of action sequences in motor learning: A connectionist approach. In Proceedings of the Ninth Annual Conference of the Cognitive Science Society (pp. 496-507). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Miyata, Y. (1988). *The learning and planning of actions* (Tech. Rep. No. 8802). San Diego, CA: University of California, Institute for Cognitive Science.

Munhall, K., & Löfqvist, A. (1987). Gestural aggregation in speech. *PAW Review, 2*, 13-15.

Munhall, K. G., & Kelso, J. A. S. (1985). Phase-dependent sensitivity to perturbation reveals the nature of speech coordinative structures [Abstract]. *Journal of the Acoustical Society of America, 78* (Suppl. 1), S38.

Munhall, K. G., Löfqvist, A., & Kelso, J. A. S. (1986). Laryngeal compensation following sudden oral perturbation [Abstract]. *Journal of the Acoustical Society of America, 80* (Suppl. 1), S109.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America, 84*, 1653-1661.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*, 151-168.

Öhman, S. E. G. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41,* 310-320.

Pearlmutter, B. A. (1988). Learning state space trajectories in recurrent neural networks. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School.* San Mateo, CA: Morgan Kaufmann.

Perkell, J. S. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press.

Rubin, P. E., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70,* 321-328.

Ruelle, D. (1980). Strange attractors. *The Mathematical Intelligencer, 2,* 126-137.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 45-76). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science, 6,* 1-36.

Saltzman, E. L. (1979). Levels of sensorimotor representation. *Journal of Mathematical Psychology, 20,* 91-163.

Saltzman, E. L. (1986). Task dynamic coordination of the speech articulators: A preliminary model. *Experimental Brain Research, Ser. 15,* 129-144.

Saltzman, E. L., Goldstein, L., Browman, C. P., & Rubin, P. (1988a). Dynamics of gestural blending during speech production [Abstract]. *Neural Networks, 1,* 316.

Saltzman, E. L., Goldstein, L., Browman, C. P., & Rubin, P. (1988b). Modeling speech production using dynamic gestural structures [Abstract]. *Journal of the Acoustical Society of America, 84* (Suppl.1), S146.

Saltzman, E. L., & Kelso, J. A. S. (1983). Toward a dynamical account of motor memory and control. In R. Magill (Ed.), *Memory and control of action* (pp. 17-38). Amsterdam: North Holland.

Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review, 94,* 84-106.

Saltzman, E. L., Rubin, P., Goldstein, L., & Browman, C. P. (1987). Task-dynamic modeling of interarticulator coordination [Abstract]. *Journal of the Acoustical Society of America, 82* (Suppl. 1), S15.

Schöner, G., Haken, H., & Kelso, J. A. S. (1986). Stochastic theory of phase transitions in human hand movement. *Biological Cybernetics, 53,* 1-11.

Scholz, J. P. (1986). *A nonequilibrium phase transition in human bimanual movement: Test of a dynamical model.* Unpublished doctoral dissertation, Department of Psychology, University of Connecticut, Storrs, CT.

Shaiman, S., & Abbs, J. H. (1987; November). *Phonetic task-specific utilization of sensorimotor actions.* Paper presented at the meeting of the American Speech-Language-Hearing Association, New Orleans, LA.

Smith, C. L., Browman, C. P., & McGowan, R. S. (1988). Applying the Program NEWPAR to extract dynamic parameters from movement trajectories [Abstract]. *Journal of the Acoustical Society of America, 84* (Suppl. 1),S128.

Stetson, R. H. (1951). *Motor phonetics: A study of speech movements in action.* Amsterdam: North Holland. Reprinted in J. A. S. Kelso & K. G. Munhall (Eds.). (1988). *R. H. Stetson's motor phonetics: A retrospective edition.* Boston: College-Hill.

Stornetta, W. S., Hogg, T., & Huberman, B. A. (1988). A dynamical approach to temporal information processing. In D. Z. Anderson (Ed.), *Neural information processing systems.* New York: American Institute of Physics.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16,* 397-420.

Tank, D. W. & Hopfield, J. J. (1987). Neural computation by concentrating information in time. *Proceedings of the National Academy of Sciences, USA, 84,* 1896-1900.

Thompson, J. M. T., & Stewart, H. B. (1986). *Nonlinear dynamics and chaos: Geometrical methods for engineers and scientists*. New York: Wiley.

Tiede, M. K. & Browman, C. P. (1988). [Articulatory x-ray and speech acoustic data for CV_1CV_2C sequences]. Unpublished research.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 211-265). Hillsdale, NJ: Lawrence Erlbaum Associates.

Turvey, M. T., Rosenblum, L. D., Schmidt, R. C., & Kugler, P. N. (1986). Fluctuations and phase symmetry in coordinated rhythmic movements. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 564-583.

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics: A cross-language study.* Bloomington, IN: Indiana University Linguistics Club.

von Holst, E. (1973). *The behavioral physiology of animal and man: The collected papers of Erich von Holst* (Vol. 1; R. Martin, trans.). London: Methuen and Co., Ltd.

Whitney, D. E. (1972). The mathematics of coordinated control of prosthetic arms and manipulators. *ASME Journal of Dynamic Systems, Measurement and Control, 94,* 303-309.

Winfree, A. T. (1980). *The geometry of biological time.* New York: Springer-Verlag.

Wing, A. M. (1980). The long and short of timing in response sequences. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 469-486). New York: North-Holland.

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics, 14,* 5-12.

Yoshioka, H., Löfqvist, A., & Collier, R. (1982). Laryngeal adjustments in Dutch voiceless obstruent production. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, 16,* 27-35.

Yoshioka, H., Löfqvist, A., & Hirose, H. (1981). Laryngeal adjustments in the production of consonant clusters and geminates in American English. *Journal of the Acoustical Society of America, 70,* 1615-1623.

FOOTNOTES

Ecological Psychology, 1989, 1(4), 333-382.

†Department of Communicative Disorders, Elborn College, University of Western Ontario, London, Ontario.

[1]The term *gesture* is used, here and elsewhere in this article, to denote a member of a family of functionally equivalent articulatory movement patterns that are *actively* controlled with reference to a given speech-relevant goal (e.g., a bilabial closure). Thus, in our usage gesture and movement have different meanings. Although gestures are composed of articulatory movements, not all movements can be interpreted as gestures or gestural components.

[2]For example, we and others have asserted that several coordinate systems (e.g., articulatory and higher-order, goal-oriented coordinates), and mappings among these coordinate systems, must be involved implicitly in the production of speech. We have adopted one method of representing these mappings explicitly in the present model (i.e., using Jacobians and Jacobian pseudoinverses; see Appendix 2). We make no strong claim, however, as to the neural or behavioral reality of these specific methods.

[3]An analogous functional partitioning has also been suggested in recent physiological studies by Lennard (1985) and Lennard and Hermanson (1985) on cyclic swimming motions of single hindlimbs in the turtle. In this work, the authors argued for a model of the locomotor neural circuit for turtle swimming that consists of two functionally distinct but interacting components. One component, analogous to the present interarticulator level, is a central intracycle pattern generator (CIPG) that organizes the patterning of muscular activity within each locomotor cycle. The second component, analogous to the present intergestural level, is an oscillatory central timing network (CTN) that is responsible for rhythmically activating or entraining the CIPG to produce an extended sequence of cycles (see also von Holst, 1973). A related distinction between "motor" and "clock" coordinative processes, respectively, has been proposed in the context of human manual rhythmic tasks consisting of either continuous oscillations at the wrist joints (e.g., Turvey, Rosenblum, Schmidt, & Kugler, 1986) or discrete finger tapping sequences (e.g., Wing, 1980; Wing & Kristofferson, 1973).

[4]We do not mean to imply that the production of vocal tract constrictions and the shaping of articulatory trajectories are the primary goals of speech production. The functional role of speech gestures is to control air pressures and flows in the vocal tract so as to produce distinctive patterns of sound. In this article, we emphasize gestural form and stability as phonetic organizing principles for the sake of relative simplicity. Ultimately, the gestural approach must come to grips with the aerodynamic sound-production requirements of speech.

[5]Since the preparation of this article, the task-dynamic model was extended to incorporate control of the tongue-tip (TTCL, TTCD), glottal (GLO), and velic (VEL) constrictions. These tract-variables and associated articulator sets are also shown in Figures 3 and 4. Results of simulations using these "new" gestures have been reported elsewhere in preliminary form (Saltzman, Goldstein, Browman, & Rubin, 1988a, 1988b).

[6]Gestural activation pulses are similar functionally to Joos's (1948) theorized "innervation waves", whose ongoing values reflected the strength of vocal tract control associated with various phonological segments or segmental components. They are also analogous to the "phonetic influence functions" used by Mattingly (1981) in the domain of acoustic speech synthesis-by-rule. Finally, the activation pulses share with Fowler's (1983) notion of segmental "prominence" the property of being related to the "extent to which vocal tract activity is given over to the production of a particular segment" (p. 392).

[7]Coarticulatory effects could also originate in two simpler ways. In the first case, "passive" coproduction could result from carryover effects associated with the cessation of active gestural control, due to the inertial sluggishness or time constants inherent in the articulatory subsystems (e.g., Coker, 1976; Henke, 1966). However, neither active nor passive coproduction need be involved in coarticulatory phenomena, at least in a theoretical sense. Even if a string of segments were produced as a temporally discrete (i.e., non-coproduced) sequence of target articulatory steady-states, coarticulatory effects on articulatory movement patterns would still result. In this second case, context-dependent differences in articulatory transitions to a given target would simply reflect corresponding differences in the interpolation of trajectories from the phonologically allowable set of immediately preceding targets. Both "sluggishness" and interpolation coarticulatory effects appear to be present in the production of actual speech.

[8]In Jordan's (1986; in press) model, a given network can learn a *single* set of weights that will allow it to produce *several* different sequences. Each such sequence is produced (and learned) in the presence of a corresponding constant activation pattern in a set of *plan* units (see Figure 10). These units provide a second set of inputs to the network's hidden layer, in addition to the inputs provided by the state units. We propose, however, to use Jordan's model for cases in which different sets of weights are learned for different sequences. In such cases, the plan units are no longer required, and we ignore them in this article for purposes of simplicity.

[9]To teach the network to perform a given sequence, Jordan (1986; in press) first initialized the network to zero, and then presented a sequence of teaching vectors (each corresponding to an element in the intended sequence), delivering one every fourth time step. At these times, errors were generated, back-propagated through the network, and the set of network weights were incrementally adjusted. During the three time steps between each teaching vector, the network was allowed to run free with no imposed teaching constraints. At the end of the teaching vector sequence the network was reinitialized to zero, and the entire weight-correction procedure was repeated until the sum of the squared output errors fell below a certain criterion. After training, the network's performance was tested starting with the state units set to zero.

[10]One possibility is to construct explicitly a set of serial mini-networks that could produce sequentially cohesive, multigesture units. Then a higher order net could be trained to produce utterance-specific sequences of such units (e.g., Jordan, 1985). It is also possible that multigesture units could arise spontaneously as emergent consequences of the learning-phase dynamics of connectionist, serial-dynamic networks that are trained to produce orchestrated patterns of the simpler gestural components (e.g., Grossberg, 1986; Miyata, 1987, 1988). This is clearly an important area to be explored in the development of our hybrid dynamical model of speech production (see the section entitled *Toward a Hybrid Dynamical Model*).

[11]Transillumination signals are uncalibrated in terms of spatial measurement scale. Consequently, amplitude differences in glottal gestures are only suggested, not demonstrated, by corresponding differences in transillumination signal size. Temporal differences (e.g., durations, glottal peak timing) and the spatiotemporal shape (e.g., one vs. two peaks) of transillumination signals are reliable indices/reflections of gestural kinematics.

[12]It is likely that rate and stress manipulations also have systematic effects on oral-glottal coordination. We make no claims regarding these potential effects in this article, however.

APPENDIX 1

Tract-variable dynamical system

The tract-variable equations of motion are defined in matrix form as follows:

$$\ddot{\mathbf{z}} = \mathbf{M}^{-1}(-\mathbf{B}\,\dot{\mathbf{z}} - \mathbf{K}\Delta\mathbf{z}), \qquad \text{(A1)}$$

where $\mathbf{z}$ = the $m \times 1$ vector of current tract-variable positions, with components z_i listed in Figure 4; $\dot{\mathbf{z}}, \ddot{\mathbf{z}}$ = the first and second derivatives of $\mathbf{z}$ with respect to time; $\mathbf{M}$ = a $m \times m$ diagonal matrix of inertial coefficients. Each diagonal element, m_{ii}, is associated with the i^{th} tract variable; $\mathbf{B}$ = a $m \times m$ diagonal matrix of tract-variable damping coefficients; $\mathbf{K}$ = a $m \times m$ diagonal matrix of tract-variable stiffness coefficients; and $\Delta\mathbf{z} = \mathbf{z} - \mathbf{z}_0$ where $\mathbf{z}_0$ = the target or rest position vector for the tract variables.

By defining the M, B, and K matrices as diagonal, the equations in (A1) are uncoupled. In this sense, the tract variables are assumed to represent independent *modes* of articulatory behavior that do not interact dynamically (see Coker, 1976, for a related use of articulatory modes). In current simulations, M is assumed to be constant and equal to the identity matrix (m_{ij} = 1.0 for $i = j$, otherwise m_{ij} = 0.0), whereas the components of B, K, and $\mathbf{z}_0$ vary during a simulated utterance according to the ongoing set of gestures being produced. For example, different vowel gestures are distinguished in part by corresponding differences in target positions for the associated set of tongue-dorsum point attractors. Similarly, vowel and consonant gestures are distinguished in part by corresponding differences in stiffness coefficients, with vowel gestures being slower (less stiff) than consonant gestures. Thus, Equation (A1) describes a linear system of tract-variable equations with time-varying coefficients, whose values are functions of the currently active gesture set (see the *Parameter Tuning* subsection of the text section *Active Gestural Control: Tuning and Gating* for a detailed account of this coefficient specification process). Note that simulations reported previously in Saltzman (1986) and Kelso et al. (1986a, 1986b) were restricted to either single "isolated" gestures, or synchronous pairs of gestures defined across different tract variables, e.g., single bilabial closures, or synchronous "vocalic" tongue-dorsum and "consonantal" bilabial gestures. In these instances, the coefficient matrices and vector parameters in Equation (A1) remained constant (time-invariant) throughout each such gesture set.

APPENDIX 2

Model articulator dynamical system; Orthogonal projection operator

A dynamical system for controlling the model articulators is specified by expressing tract variables ($\mathbf{z}$, $\dot{\mathbf{z}}$, $\ddot{\mathbf{z}}$) as functions of the corresponding model articulator variables ($\boldsymbol{\phi}$, $\dot{\boldsymbol{\phi}}$, $\ddot{\boldsymbol{\phi}}$). The tract variables of Equation (A1) are transformed into model articulator variables using the following direct kinematic relationships:

$$\mathbf{z} = \mathbf{z}(\boldsymbol{\phi}) \qquad \text{(A2a)}$$

$$\dot{\mathbf{z}} = \mathbf{J}(\boldsymbol{\phi})\,\dot{\boldsymbol{\phi}} \qquad \text{(A2b)}$$

$$\ddot{\mathbf{z}} = \mathbf{J}(\boldsymbol{\phi})\,\ddot{\boldsymbol{\phi}} + \dot{\mathbf{J}}(\boldsymbol{\phi}, \dot{\boldsymbol{\phi}})\dot{\boldsymbol{\phi}}, \qquad \text{(A2c)}$$

where $\boldsymbol{\phi}$ = the $n \times 1$ vector of current articulator positions, with components ϕ_j listed in Figure 4; $\mathbf{z}(\boldsymbol{\phi})$ = the current $m \times 1$ tract-variable position vector expressed as a function of the current model articulator configuration. These functions are specific to the particular geometry assumed for the set of model articulators used to simulate speech gestures or produce speech acoustics via articulatory synthesis. $\mathbf{J}(\boldsymbol{\phi})$ = the $m \times n$ *Jacobian* transformation matrix whose elements J_{ij} are partial derivatives, $\partial z_i/\partial \phi_j$, evaluated at the current $\boldsymbol{\phi}$. Thus, each row-i of the Jacobian represents the set of changes in the i^{th} tract variable resulting from unit changes in all the articulators; and $\dot{\mathbf{J}}(\boldsymbol{\phi}, \dot{\boldsymbol{\phi}}) = (d\mathbf{J}(\boldsymbol{\phi})/dt)$, a $m \times n$ matrix resulting from differentiating the elements of $\mathbf{J}(\boldsymbol{\phi})$ with respect to time. The elements of $\dot{\mathbf{J}}$ are functions of both the current $\boldsymbol{\phi}$ and $\dot{\boldsymbol{\phi}}$. The elements of J and $\dot{\mathbf{J}}$ thus reflect the geometrical relationships among motions of the model articulators and motions of the corresponding tract variables. Using the direct kinematic relationships in Equation (A2), the equation of motion derived for the actively controlled model articulators is as follows:

$$\ddot{\boldsymbol{\phi}}_A = \mathbf{J}^*(\mathbf{M}^{-1}[-\mathbf{B}\mathbf{J}\,\dot{\boldsymbol{\phi}} - \mathbf{K}\Delta\mathbf{z}(\boldsymbol{\phi})]) - \mathbf{J}^*\,\dot{\mathbf{J}}\dot{\boldsymbol{\phi}}, \qquad \text{(A3)}$$

where $\boldsymbol{\phi}_A$ = an articulatory acceleration vector representing the active driving influences on the model articulators; M, B, K, J, and $\dot{\mathbf{J}}$ are the same matrices used in Equations (A1) and (A2); $\Delta\mathbf{z}(\boldsymbol{\phi}) = \mathbf{z}(\boldsymbol{\phi}) - \mathbf{z}_0$, where $\mathbf{z}_0$ = the same constant vector used in Equation (A1); It should be noted that because $\Delta\mathbf{z}$ in Equations (A1) and (A3) is *not* assumed to be "small," a differential approximation $d\mathbf{z} = \mathbf{J}(\boldsymbol{\phi})d\boldsymbol{\phi}$ is not justified and,

therefore, Equation (A2a) was used instead for the kinematic displacement transformation into model articulator variables; J^* = a $n \times m$ *weighted Jacobian pseudoinverse* (e.g., Benati, Gaglio, Morasso, Tagliasco, & Zaccaria, 1980; Klein & Huang, 1983; Whitney, 1972). $J^* = W^{-1}J^T(JW^{-1}J^T)^{-1}$ where W is a $n \times n$ positive definite *articulatory weighting* matrix whose elements are constant during a given isolated gesture, and superscript T denotes the vector or matrix *transpose* operation. The pseudoinverse is used because there are a greater number of model articulator variables than tract variables for this task. More specifically, using J^* provides a unique, optimal least squares solution for the *redundant* (e.g., Saltzman, 1979) differential transformation from tract variables to model articulator variables that is weighted according to the pattern of elements in the W-matrix. In current modeling, the W-matrix is defined to be of diagonal form, in which element w_{jj} is associated with articulator ϖ_j. A given set of articulator weights implements a corresponding pattern of constraints on the relative motions of the articulators during a given gesture. The motion of a given articulator is constrained in direct proportion to the magnitude of the corresponding weighting element relative to the remaining weighting elements. Intuitively, then, the elements of W establish a gesture-specific pattern of relative "receptivities" among the articulators to the driving influences generated in the tract-variable state space. In the present model, J^* has been generalized to a form whose elements are gated functions of the currently active gesture set (see the *Transformation gating* subsection of the text section *Active gestural control: Tuning and gating* for details).

In Equation (A3), the first and second terms inside the inner parentheses on the right hand side represent the articulatory acceleration components due to system damping ($\ddot{\varpi}_d$) and stiffness ($\ddot{\varpi}_s$), respectively. The rightmost term on the right hand side represents an acceleration component vector ($\ddot{\varpi}_{vp}$) that is nonlinearly proportional to the squares and pairwise products of current articulatory velocities (e.g., $(\dot{\varpi}_2)^2$, $\dot{\varpi}_2\dot{\varpi}_3$, etc.; for further details, see Kelso et al., 1986a, 1986b; Saltzman, 1986; Saltzman & Kelso, 1987).

In early simulations of unperturbed discrete speech gestures (e.g., bilabial closure) it was found that, after a given gestural target (e.g., degree of lip compression) was attained and maintained at a steady value, the articulators continued to move with very small but non-negligible (and undesirable) velocities. In essence, the model added to the articulator movements just those patterns that resulted in no tract-variable (e.g., lip aperture) motion above and beyond that demanded by the task. The source of this residual motion was ascertained to reside in the nonconservative nature of the pseudoinverse (J^*; see Equation [A3]) of the Jacobian transformation (J) used to relate tract-variable motions and model articulator motions (Klein & Huang, 1983). By nonconservative, we mean that a closed path in tract-variable space does not imply generally a closed path in model articulator space.

These undesired extraneous model-articulator motions were eliminated by including supplementary dissipative forces proportional to the articulatory velocities. Specifically, the *orthogonal projection operator*, $(I_n - J^*J])$, where I_n is a $n \times n$ identity matrix (Ballieul, Hollerbach & Brocket, 1984; Klein & Huang, 1983) was used in the following augmented form of Equation (A3):

$$\ddot{\varpi}_A = J^*(M^{-1}[- BJ\dot{\varpi} - K\Delta z(\varpi)]) - J^*\dot{J}\dot{\varpi} + (I_n - J^*J))\ddot{\varpi}_d, \quad \text{(A4)}$$

where $\ddot{\varpi}_d = B_N\dot{\varpi}$ represents an acceleration damping vector, and B_N is a $n \times n$ diagonal matrix whose components, b_{Nij}, serve as constant damping coefficients for the j^{th} component of $\ddot{\varpi}$. The subscript N denotes the fact that B_N is the same damping matrix as that used in the articulatory *neutral* attractor (see the text section on *Nonactive Gestural Control*, Equations [4] and [5]).

Using Equation (A4), the model generates movements to tract-variable targets with no residual motions in either tract-variable or model-articulator coordinates. Significantly, the model works equally well for both the case of unperturbed gestures, and the case in which gestures are perturbed by simulated external mechanical forces (see the text section *Gestural primitives*). In the present model, the identity matrix (I_n) in Equation (A4) has been generalized, like J^*, to a form whose elements are gated functions of the currently active gesture set (see the *Transformation gating* subsection of the text section *Active gestural control: Tuning and gating*).

The damping coefficients of B_N are typically assigned equal values for all articulators. This results in synchronous movements (varying in

amplitudes) for the tract variables and articulators involved in isolated gestures. Interesting patterns emerge, however, if the coefficients are assumed to be unequal for the various articulators (Saltzman et al., 1987). For example, the relatively sluggish rotations of the jaw or horizontal motions of the lips may be characterized by larger time constants than the lips' relatively brisk vertical motions. Implementing these asymmetries into B_N, interarticulator asynchronies within single speech gestures are generated by the model that mirror, partially, some patterns reported in the literature. For example, Gracco and Abbs (1986) showed that, during bilabial closing gestures for the first /p/ in /sæpæpl/, the raising onsets and peak velocities of the component articulatory movements occur in the order: upper lip, lower lip, and jaw. The peak velocities conform to this order more closely than the raising onsets. In current simulations of isolated bilabial gestures, the asynchronous pattern of the peak velocities (but not the movement onsets) emerges naturally when the elements of B_N are unequal. Interestingly, the tract-variable trajectories are identical to those generated when B_N's elements are equal. Additional simulations have revealed that patterns of closing onsets may become asynchronous, however, depending on several factors, e.g., the direction and magnitude of the jaw's velocity prior to the onset of the closing gesture.

APPENDIX 3

Competitive network equations for parameter tuning

The *post*blending activation strengths (p_{Tik} and p_{Wikj}) defined in text Equation (2) are given by the steady-state solutions to a set of feedforward, competitive-interaction-network dynamical equations (e.g., Grossberg, 1986) for the preblending activation strengths (a_{ik}) in the present model. These equations are expressed as follows:

$$p_{Tik} = -a_{ik}\left(p_{Tik} - B_p\right) - \left(\left[B_a - a_{ik}\right] + \beta_{ik} \sum_{\substack{l \in Z_i \\ l \neq k}} \left[\alpha_{il} a_{il}\right]\right) p_{Tik} \text{ , and} \tag{A5a}$$

$$p_{Wikj} = -a_{ik}\left(p_{Wikj} - B_p\right) - \left(\left[B_a - a_{ik}\right] + \beta_{ik} \sum_{i \in \Phi_j} \left[\sum_{\substack{l \in Z_i \\ l \neq k}} a_{il} a_{il}\right]\right) p_{Wikj} \text{ ,} \tag{A5b}$$

where B_p and B_a denote the maximum values allowed for the pre-blending and post-blending activation strengths, respectively. In current modeling, B_p and B_a are defined to equal 1.0; and a_{il} and β_{ik} are the lateral inhibition and "gatekeeper" coefficients, respectively, defined in text Equation (2).

The solutions to Equations (A5a) and (A5b) are obtained by setting their left-hand sides to zero, and solving for p_{Tik} and p_{Wikj}, respectively. These solutions are expressed in Equations (2a) and (2b). The dynamics of Equation (A5) are assumed to be "fast" relative to the dynamics of the interarticulator coordination level (Equations [A3] and [A4]). Consequently, incorporating the solutions of Equation (A5) directly into Equation (1) is viewed as a justified computational convenience in the present model (see also Grossberg & Mingolla, 1986, for a similar computational simplification).

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 69-101

Articulatory Gestures as Phonological Units*

Catherine P. Browman and Louis Goldstein†

1 A GESTURAL PHONOLOGY

Over the past few years, we have been investigating a particular hypothesis about the nature of the basic 'atoms' out of which phonological structures are formed. The atoms are assumed to be primitive actions of the vocal tract articulators that we call 'gestures.' Informally, a gesture is identified with the formation (and release) of a characteristic constriction within one of the relatively independent articulatory subsystems of the vocal tract (i.e., oral, laryngeal, velic). Within the oral subsystem, constrictions can be formed by the action of one of three relatively independent sets of articulators: the lips, the tongue tip/blade and the tongue body. As actions, gestures have some intrinsic time associated with them—they are characterizations of movements through space and over time (see Fowler et al., 1980).

Within the view we are developing, phonological structures are stable 'constellations' (or 'molecules', to avoid mixing metaphors) assembled out of these gestural atoms. In this paper, we examine some of the evidence for, and some of the consequences of, the assumption that gestures are the basic atoms of phonological structures. First, we attempt to establish that gestures are pre-linguistic discrete units of action that are inherent in the maturation of a developing child and that therefore can be harnessed as elements of a phonological system in the course of development (§ 1.1). We then give a more detailed, formal characterization of gestures as phonological units within the context of a computational model (§ 1.2), and show that a number of phonological regularities can be captured by representing constellations of gestures (each having inherent duration) using gestural scores (§ 1.3). Finally, we show how the proposed gestural structures relate to proposals of feature geometry (§§ 2 - 3).

This paper has benefited from criticisms by Cathi Best, Alice Faber, Elliot Saltzman, Michael Studdert-Kennedy, Eric Vatikiotis-Bateson, Doug Whalen and two anonymous reviewers. Our thanks to Mark Tiede for help with manuscript preparation, and Zefang Wang for help with the graphics. This work was supported by NSF grant BNS-8520709 and NIH grants HD-01994 and NS-13617 to Haskins Laboratories.

1.1 Gestures as pre-linguistic primitives

Gestures are units of action that can be identified by observing the coordinated movements of vocal tract articulators. That is, repeated observations of the production of a given utterance will reveal a characteristic pattern of constrictions being formed and released. The fact that these patterns of (discrete) gestures are similar in structure to the nonlinear phonological representations being currently postulated (e.g., Clements, 1985; Hayes, 1986; Sagey, 1986), together with some of the evidence presented in Browman and Goldstein (1986, in press), leads us to make the strong hypothesis that gestures themselves constitute basic phonological units. This hypothesis has the attractive feature that the basic units of phonology can be identified directly with cohesive patterns of movement within the vocal tract. Thus, the phonological system is built out of inherently discrete units of action. This state of affairs would be particularly useful for a child learning to speak. If we assume that discrete gestures (like those that will eventually function as phonological units) emerge in the child's behavioral repertoire *in advance* of any specifically linguistic development, then it is possible to view phonological development as harnessing these action units to be the basic units of phonological structures.

The idea that pre-linguistic gestures are employed in the service of producing early words has been proposed and supported by a number of writers, for example, Fry (1966, in Vihman in press), Locke (1983), Studdert-Kennedy (1987) and Vihman (in press), where what we identify as 'gestures' are referred to as 'articulatory routines' or the like. The view we are proposing extends

this approach by hypothesizing that these pre-linguistic gestures actually become the units of contrast. Additional phonological developments involve differentiating and tuning the gestures, and developing patterns of intergestural coordination that correspond to larger phonological structures.

The evidence that gestures are pre-linguistic units of action can be seen in the babbling behavior of young infants. The descriptions of infant babbling (ages 6-12 months) suggest a predominance of what are transcribed as simple CV syllables (Locke, 1986; Oller & Eilers, 1982). The 'consonantal' part of these productions can be analyzed as simple, gross, constriction maneuvers of the independent vocal tract subsystems and (within the oral subsystem) the separate oral articulator sets. For example, based on frequency counts obtained from a number of studies, Locke (1983) finds that the consonants in (1) constitute the 'core' babbling inventory: these 12 'consonants' account for about 95% of the babbles of English babies. Similar frequencies obtain in other language environments.

(1) h bdg ptk mn jw s

These transcriptions are not meant to be either systematic phonological representations (the child doesn't have a phonology yet), or narrow phonetic transcriptions (the child cannot be producing the detailed units of its 'target' language, because, as noted below, there do not seem to be systematic differences in the babbles produced by infants in different language environments). Others have noted the problems inherent in using a transcription that assumes a system of units and relations to describe a behavior that lacks such a system (e.g., Kent & Murray, 1982; Koopmans-van Beinum & van der Stelt, 1986; Oller, 1986; Studdert-Kennedy, 1987). As Studdert-Kennedy (1987) argues, it seems likely that these transcriptions reflect the production by the infant of simple vocal constriction gestures, of the kind that evolve into mature phonological structures (which is why adults can transcribe them using their phonological categories). Thus, /h/ can be interpreted as a laryngeal widening gesture and /bdg/ as 'gross' constriction gestures of the three independent oral articulator sets (lips, tongue tip, and tongue body). /ptk/ combine the oral constriction gestures with the laryngeal maneuver, and /m n/ combine oral constrictions with velic lowering. These combinations do not necessarily indicate an ability on the part of the infant to coordinate the gestures. Rather, any accidental temporal coincidence of two such gestures would be perceived by the listener as the segments in question.

The analysis outlined above suggests that babbling involves the emergence, in the infant, of simple constriction gestures of independent parts of the vocal tract. As argued by Locke (1986), the pattern of emergence of these actions can be viewed as a function of anatomical and neurophysiological developments, rather than the beginning of language acquisition, *per se*. This can be seen, first of all, in the fact that the babbling inventory and its developmental sequence have not been shown to vary as a function of the particular language environment in which the child finds itself (although individual infants may vary considerably from one another in the relative frequencies of particular gestures—Studdert-Kennedy 1987; Vihman, in press). In fact, in the large number of studies reviewed by Locke (1983), there appear to be no detectable differences (either instrumentally or perceptually) in the 'consonantal' babbling of infants reared in different language environments. (More recent studies have found some language environment effect on the overall long term spectrum of vocalic utterances—de Boysson-Bardies et al., 1986; and on prosody—de Boysson-Bardies et al., 1984. Other subtle effects may be uncovered with improvement of analytic techniques.)

Secondly, Locke (1983) notes that the developmental changes in frequency of particular babbled consonants can likely be explained by anatomical developments. Most of the consonants produced by very young infants (less than six months) involve tongue body constrictions, usually transcribed as velars. Some time shortly after the beginning of repetitive canonical babbling (usually in the seventh month), tongue tip and lip constrictions begin to outnumber tongue body constrictions, with tongue tip constrictions eventually dominating. Even deaf infants show a progression that is qualitatively similar, at least in early stages, although their babbling can be distinguished from that of hearing infants on a number of acoustic measures (Oller & Eilers, 1988). Locke suggests an explanation in terms of vocal tract maturation. At birth, the infant's larynx is high, and the tongue virtually fills the oral cavity (Lieberman, 1984). This would account for the early dominance of tongue body constrictions. After the larynx drops, tongue tip and lip constrictions—without simultaneous tongue constrictions—are more readily formed. In particular, the closing action of the mandible will

then contribute to constrictions at the front of the mouth.

Finally, Locke (1986) notes that the timing of the development of repetitive 'syllabic' babbling coincides with the emergence of repetitive motor behaviors generally. He cites Thelen's (1981) observation of 47 different rhythmic activities that have their peak frequency at 6-7 months. Locke concludes (1986, p. 145) that 'it thus appears that the syllabic patterning of babble—like the phonetic patterning of its segments—is determined mostly by nonlinguistic developments of vocal tract anatomy and neurophysiology.'

The pre-linguistic vocal gestures become linguistically significant when the child begins to produce its first few words. The child seems to notice the similarity of its babbled patterns to the speech s/he hears (Locke 1986), and begins to produce 'words' (with apparent referential meaning) using the available set of vocal gestures. It is possible to establish that there is a definite relationship between the (nonlinguistic) gestures of babbling and the gestures employed in early words by examining individual differences among children. Vihman et al. (1985) and Vihman (in press) find that the particular consonants that were produced with high frequency in the babbling of a given child also appear with high frequency in that child's early word productions. Thus, the child is recruiting its well-practiced action units for a new task. In fact, in some early cases (e.g., 'baby' words like *mama*, etc.), 'recruiting' is too active a notion. Rather, parents are helping the child establish a referential function with sequences that already exist as part of the babbling repertoire (Locke, 1986).

Once the child begins producing words (complex units that have to be distinguished one from another) using the available gestures as building blocks, phonology has begun to form. If we compare the child's early productions (using the small set of pre-linguistic gestures) to the gestural structure of the adult forms, it is clear that there are (at least) two important developments that are required to get from one to the other: (1) differentiation and tuning of individual gestures and (2) coordination of the individual gestures belonging to a given word. Let us examine these in turn.

Differentiation and tuning. While the repertoire of gestures inherent in the consonants of (1) above employs all of the relatively independent articulator sets, the babbled gestures involve just a single (presumably gross) movement. For example, some kind of closure is involved for oral constriction gestures. In general, however, languages employ gestures produced with a given articulator set that contrast in the degree of constriction. That is, not only are closure gestures produced but also fricative and wider (approximant) gestures. In addition, the exact location of the constriction formed by a given articulator set may contrast, e.g., in the gestures for /θ/, /s/ and /ʃ/. Thus, a single pre-linguistic constriction gesture must eventually differentiate into a variety of potentially contrastive gestures, tuned with different values of constriction location and degree. Although the location and degree of the constriction formed by a given articulator set are, in principle, physical continua, the differentiated gestures can be categorically distinct. The partitioning of these continua into discrete categories is likely aided by quantal (i.e., nonlinear) articulatory-auditory relations of the kind proposed by Stevens (1972, 1989). In addition, Lindblom (1986) has shown how the pressures to keep contrasting words perceptually distinct can lead to discrete clustering along some articulatory/acoustic contina. Even so, the process of differentiation may be lengthy. Nittrouer, Studdert-Kennedy, and McGowan (1989) present data on fricative production in American English children. They find that differentiation between /s/ and /ʃ/ is increasing in children from ages three to seven, and hasn't yet reached the level shown by adults. In addition, tuning may occur even where differentiation is unnecessary. That is, even if a language has only a single tongue tip closure gesture, its constriction location may be tuned to a particular language-specific value. For example, English stops have an alveolar constriction location, while French stops are more typically dental.

Coordination. The various gestures that constitute the atoms of a given word must be organized appropriately. There is some evidence that a child can know what all the relevant gestures are for some particular word, and can produce them all, without either knowing or being able to produce the appropriate organization. Studdert-Kennedy (1987) presents an example of this kind from Ferguson and Farwell (1975). They list ten attempts by a 15-month old girl to say the word *pen* in a half-hour session, as shown in (2):

(2) [mə̃^{ə}, ^{n}ɪ̃, de^{dn}, hɪn, ^{m}bə̃, p^{h}ɪn, t^{h}ɳt^{h}ɳt^{h}ɳ, ba^{h}, ɖ^{h}au^{n}, buə̃]

While these attempts appear radically different, they can be analyzed, for the most part, as the set

of gestures that constitute *pen* misarranged in various ways: glottal opening, bilabial closure, tongue body lowering, alveolar closure and velum lowering. Eventually, the 'right' organization is hit upon by the child. The search is presumably aided by the fact that the coordinated structure embodied in the target language is one of a relatively small number of dynamically stable patterns (see also Boucher, 1988). The formation of such stable patterns may ultimately be illuminated by research into stable modes in coordinated human action in general (e.g., Haken et al., 1985; Schmidt et al., 1987) being conducted within the broad context of the non-linear dynamics relevant to problems of pattern formation in physics and biology (e.g., Glass & Mackey, 1988; Thompson & Stewart, 1976). In addition, aspects of coordination may emerge as the result of keeping a growing number of words perceptually distinct using limited articulatory resources (Lindblom et al., 1983).

If phonological structures are assumed to be organized patterns of gestural units, a distinct methodological bonus obtains: the vocal behavior of infants, even pre-linguistic behavior, can be described using the same primitives (discrete units of vocal action) that are used to describe the fully elaborated phonological system of adults. This allows the growth of phonological form to be precisely monitored, by observing the development of the primitive gestural structures of infants into the elaborated structures of adults (Best & Wilkenfeld, 1988 provide an example of this). In addition, some of the thorny problems associated with the transcription of babbling can be obviated. Discrete units of action are present in the infant, and can be so represented, even if adult-like phonological structures have not yet developed. A similar advantage applies to describing various kinds of 'disordered' speech (e.g., Kent, 1983; Marshall et al., 1988), which may lack the organization shown in 'normal' adult phonology, making conventional phonological/phonetic transcriptions inappropriate, but which may, nevertheless, be composed of gestural primitives. Of course, all this assumes that it is possible to give an account of adult phonology using gestures as the basic units—it is to that account that we now turn.

1.2 The nature of phonological gestures

In conjunction with our colleagues Elliot Saltzman and Philip Rubin at Haskins Laboratories, we are developing a computational model that produces speech beginning with a representation of phonological structures in terms of gestures (Browman et al., 1984; Browman et al., 1986; Browman & Goldstein, 1987; Saltzman et al., 1987), where a *gesture* is an abstract characterization of coordinated task-directed movements of articulators within the vocal tract. Each gesture is precisely defined in terms of the parameters of a set of equations for a 'task-dynamic' model (Saltzman, 1986; Saltzman & Kelso, 1987). When the control regime for a given gesture is active, the equations regulate the coordination of the model's articulators in such a way that the gestural 'task' (the formation of a specified constriction) is reached as the articulator motions unfold over time. Acoustic output is obtained from these articulator motions by means of an articulatory synthesizer (Rubin et al., 1981). The gestures for a given utterance are themselves organized into a larger coordinated structure, or constellation, that is represented in a *gestural score* (discussed in § 1.3). The score specifies the sets of values of the dynamic parameters for each gesture, and the temporal intervals during which each gesture is active. While we use analyses of articulatory movement data to determine the parameter values for the gestures and gestural scores, there is nevertheless a striking convergence between the structures we derive through these analyses and phonological structures currently being proposed in other frameworks (e.g., Anderson & Ewen, 1987; Clements, 1985; Ewen, 1982; Lass, 1984; McCarthy, 1989; Plotkin, 1976; Sagey, 1986).

Within task dynamics, the goal for a given gesture is specified in terms of independent task dimensions, called *vocal tract variables*. Each tract variable is associated with the specific sets of articulators whose movements determine the value of that variable. For example, one such tract variable is Lip Aperture (LA), corresponding to the vertical distance between the two lips. Three articulators can contribute to changing LA: the jaw, vertical displacement of the lower lip with respect to the jaw, and vertical displacement of the upper lip. The current set of tract variables in the computational model, and their associated articulators, can be seen in Figure 1. Within the task dynamic model, the control regime for a given gesture coordinates the ongoing movements of these articulators in a flexible, but task-specific manner, according to the demands of other concurrently active gestures. The motion associated with each of a gesture's tract variables is specified in terms of an equation for a second-order dynamical system.[1] The equilibrium

position parameter of the equation [x0] specifies the tract variable target that will be achieved, and the stiffness parameter *k* specifies (roughly) the time required to get to target. These parameters are tuned differently for different gestures. In addition, their values can be modified by stress.

Gestures are currently specified in terms of one or two tract variables. Velic gestures involve a single tract variable of aperture size, as do glottal gestures. Oral gestures involve pairs of tract variables that specify the constriction degree (LA, TTCD, and TBCD) and constriction location (LP, TTCL, and TBCL). For simplicity, we will refer to the sets of articulators involved in oral gestures using the name of the end-effector, that is, the name of the single articulator at the end of the chain of articulators forming the particular constriction: the LIPS for LA and LP, the tongue tip (TT) for gestures involving TTCD and TTCL, and the tongue body (TB) for gestures involving TBCD and TBCL. As noted above, each tract variable is modelled using a separate dynamical equation; however, at present the paired tract variables use identical stiffness and are activated and de-activated simultaneously. The damping parameter *b* for oral gestures is always set for critical damping—the gestures approach their targets, but do not overshoot it, or 'ring.' Thus, a given oral gesture is specified by the values of three parameters: target values for each of a pair of tract variables, and a stiffness value (used for both equations).

This set of tract variables is not yet complete, of course. Other oral tract variables that need to be implemented include an independent tongue root variable (Ladefoged & Halle, 1988), and (as discussed in § 2.1) variables for controlling the shape of TT and TB constrictions as seen in the third dimension. Additional laryngeal variables are required to allow for pitch control and for vertical movement of the larynx, required, for example, for ejectives and implosives.

	tract variable	articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, body, jaw
TTCD	tongue tip constrict degree	tongue tip, body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

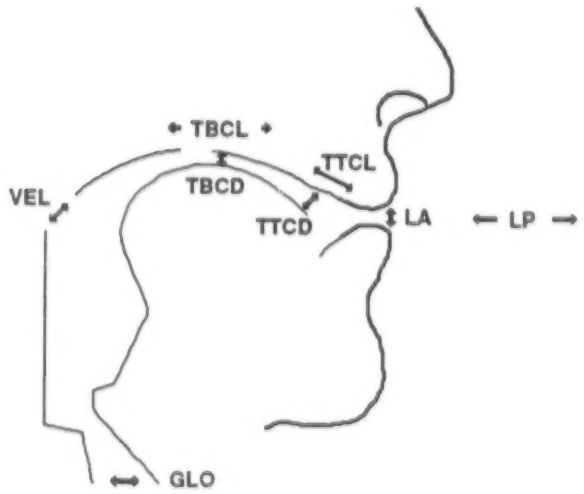


Figure 1. Tract variables and contributing articulators of computational model.

The representations of gestures employing distinct sets of tract variables are categorically distinct within the system outlined here. That is, they are defined using different variables that correspond to different sets of articulators. They provide, therefore, an inherent basis for contrast among gestures (Browman & Goldstein, in press). However, for contrasting gestures that employ the same tract variables, the difference between the gestures is in the tuned values of the continuous dynamic parameters (for oral gestures: constriction degree, location and stiffness). That is, unlike the articulator sets being used, the dynamic parameters do not inherently define categorically distinct classes. Nonetheless, we assume that there are stable ranges of parameter values that tend to contrast with one another repeatedly in languages (Ladefoged & Maddieson, 1986; Vatikiotis-Bateson, 1988). The discrete values might be derived using a combination of articulatory and auditory constraints applied across the entire lexicon of a language, as proposed by Lindblom et al. (1983). In addition, part of the basis for the different ranges might reside in the nonlinear relation between the parameter values and their acoustic consequences, as in Stevens' quantal theory (Stevens 1972, 1989).In order to represent the contrastive ranges of gestural parameter values in a discrete fashion, we employ a set of *gestural descriptors*. These descriptors serve as pointers to the particular articulator set involved in a given gesture, and to the numerical values of the dynamical parameters characterizing the gestures. In addition, they can act as classificatory and distinctive features for the purposes of lexical and phonological structure. Every gesture can be specified by a distinct descriptor structure. This functional organization can be formally represented as in (3), which relates the parameters of the dynamical equations to the symbolic descriptors. Contrasting gestures will differ in at least one of these descriptors.

(3) Gesture = articulator set (constriction degree, constriction location, constriction shape, stiffness)

Constriction Degree is always present, and refers to the $x0$ value for the constriction degree tract variables (LA, TTCD, TBCD, VEL, or GLO).

Constriction Location is relevant only for oral gestures, and refers to the $x0$ value for the constriction location tract variables (LP, TTCL, or TBCL).

Constriction Shape is relevant only for oral gestures, and refers to the $x0$ value of constriction shape tract variables. It is not currently implemented.

Stiffness refers to the k value of the tract variables.

Figure 2 displays the inventory of articulator sets and associated parameters that we posit are required for a general gestural phonology. The parameters correspond to the particular tract variables of the model shown below them. Those parameters with asterisks are not currently implemented.

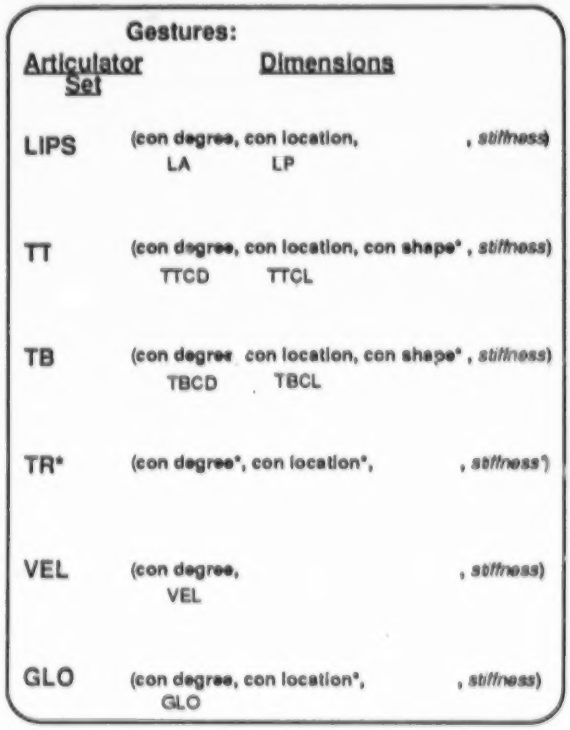


Figure 2. Inventory of articulator sets and associated parameters.

For the present, we list without comment the possible descriptor values for the constriction degree (CD) and constriction location (CL) dimensions in (4). In § 2.1, we will discuss the gestural dimensions and these descriptors in detail, including a comparison to current proposals of featural geometry.

(4) CD descriptors: closed critical narrow mid wide
CL descriptors: protruded labial dental alveolar post-alveolar palatal velar uvular pharyngeal

In the phonology of dynamically defined articulatory gestures that we are developing, gestures are posited to be the atoms of phonological structure. It is important to note that such gestures are relatively abstract. That is, the physically continuous movement trajectories are analyzed as resulting from a set of discrete, concurrently active gestural control regimes. They are discrete in two senses: (1) the dynamic parameters of a gesture's control regime remain constant throughout the discrete interval of time during which the gesture is active, and (2) gestures in a language may differ from one another in discrete ways, as represented by different descriptor values. Thus, as argued in Browman and Goldstein (1986) and Browman and Goldstein (in press), the gestures for a given utterance, together with their temporal patterning, perform a dual function. They characterize the actual observed articulator movements (thus obviating the need for any additional implementation rules), and they also function as units of contrast (and more generally capture aspects of phonological patterning). As discussed in those papers, the gesture as a phonological unit differs both from the feature and from the segment (or root node in current feature geometries). It is a larger unit than the feature, being effectively a unitary constriction action, parameterized jointly by a linked structure of features (descriptor values). Yet it is a smaller unit than the segment: several gestures linked together are necessary to form a unit at the segmental, or higher, levels.

1.3 Gestural scores: Articulatory tiers and internal duration

In the preceding section, gestures were defined with reference to a dynamical system that shapes patterns of articulatory movements. Each gesture possesses, therefore, not only an inherent spatial aspect (i.e., a tract variable goal) but also an intrinsic temporal aspect (i.e., a gestural stiffness). Much of the power of the gestural approach follows from these basic facts about gestures (combined with their abstractness), since they allow gestures to *overlap* in time as well as in articulator and/or tract variable space (see also Bell-Berti & Harris, 1981; Fowler, 1980, 1983; Fujimura, 1981a,b). In this section, we show how overlap among gestures is represented, and demonstrate that simple changes in the patterns of overlap between neighboring gestural units can automatically produce a variety of superficially different types of phonetic and phonological variation.

Within the computational model described above, the pattern of organization, or *constellation*, of gestures corresponding to a given utterance is embodied in a set of phasing principles (see Kelso & Tuller, 1987; Nittrouer et al., 1988) that specify the spatiotemporal coordination of the gestures (Browman & Goldstein, 1986). The pattern of intergestural coordination that results from applying the phasing principles, along with the interval of active control for individual gestures, is displayed in a two-dimensional *gestural score*, with articulatory tiers on one dimension and temporal information on the other (Browman et al., 1986; Browman & Goldstein, 1987). A gestural score for the word *palm* (pronounced [pɑm]) is displayed in Figure 3a. As can be seen in the figure, the tiers in a gestural score, on the vertical axis, represent the sets of articulators (or the relevant subset thereof) employed by the gestures, while the horizontal dimension codes time.

The boxes in Figure 3a correspond to individual gestures, labelled by their descriptor values for constriction degree and constriction location (where relevant). For example, the initial oral gesture is a bilabial closure, represented as a constriction of the LIPS, with a [closed] constriction degree, and a [labial] constriction location. The horizontal extent of each box represents the interval of time during which that particular gesture is active. During these activation intervals, which are determined in the computational model from the phasing principles and the inherent stiffnesses of each gesture, the particular set of dynamic parameter values that defines each gesture is actively contributing to shaping the movements of the model articulators. In Figure 3b, curves are added that show the time-varying tract variable trajectories generated by the task dynamic model according to the parameters indicated by the boxes. For example, during the activation interval of the initial labial closure, the curve representing LA (the vertical distance between the lips) decreases. As can be seen in these curves, the activation intervals directly capture something about the durations of the movements of the gestures.

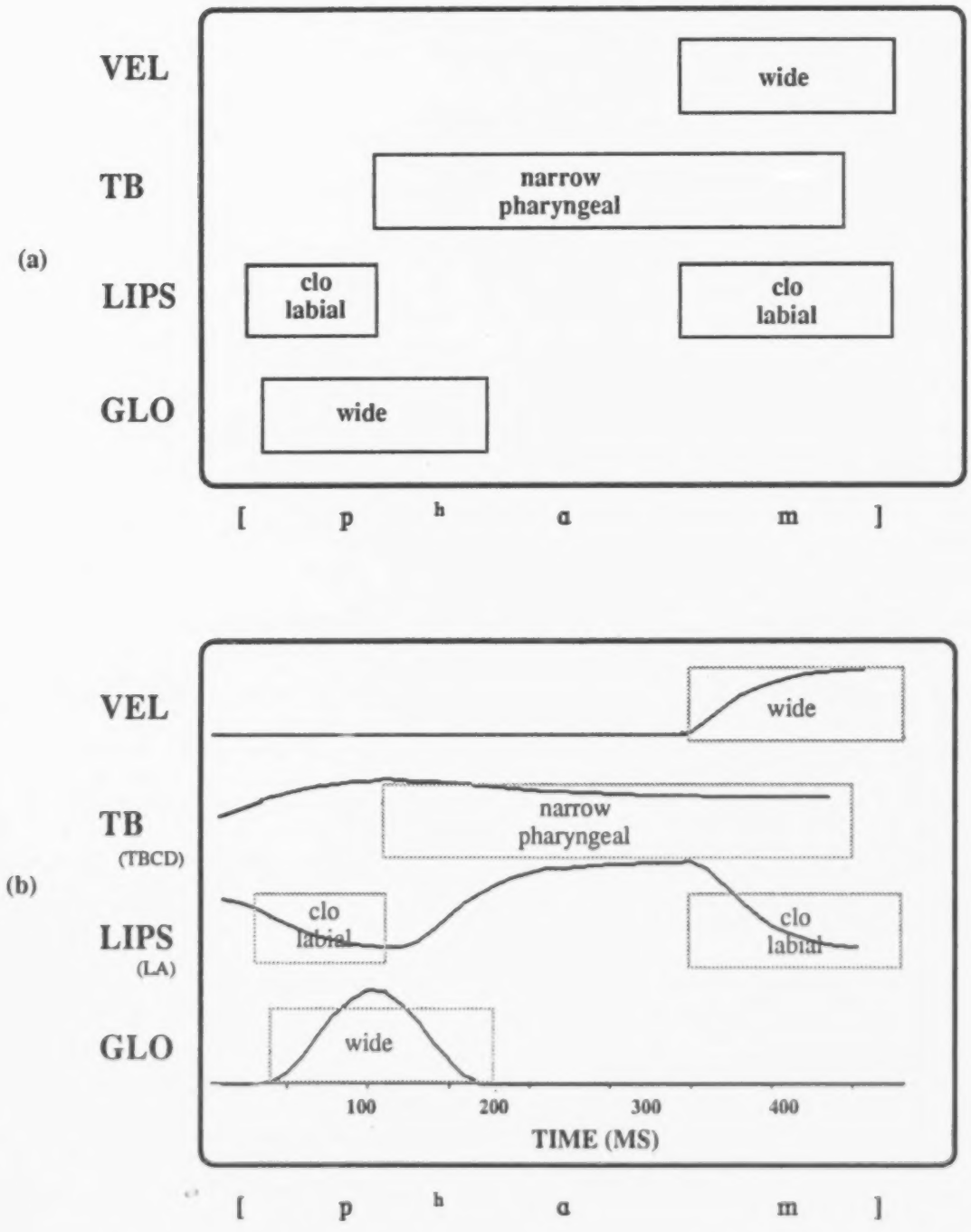


Figure 3. **Gestural score for *palm* [pɑm] using box notation. *(a)* Activation intervals only; *(b)* model generated tract variable motions added.**

Figure 4 presents an alternative symbolic redisplay[2] of the gestural score. Instead of the 'box' notation of Figure 3, a 'point' notation is employed that references only the gestural descriptors and relative timing of their 'targets.' The extent of activation of the gestures in the box notation is not indicated. In addition, association lines between the gestures have been added. These lines indicate which gestures are phased with respect to each other. Thus, the display is a shorthand representation of the phasing rules discussed in Browman and Goldstein (1987). The pair of gestures connected by a given association line are coordinated so that a specified phase of one gesture is synchronized with some phase of the other. For example, the peak opening of the GLO [wide] gesture (180 degrees) is synchronized with the release phase (290 degrees) of the LIPS [clo labial] gesture. Also important for the phase rules is the projection of oral constriction gestures onto separate Vowel and Consonant tiers, which are not shown here (Browman & Goldstein, 1987, 1988; see also Keating, 1985).

The use of point notation highlights the association lines, and, as we shall see in § 2.2.2, is useful for the purpose of comparing gestural organizations with feature geometry representations in which individual units lack any extent in time. For the remainder of this section, however, we will be concerned with showing the extent of temporal overlap of gestures, and therefore will employ the box notation form of the gestural score.

The information represented in the gestural score serves to identify a particular lexical entry. The basic elements in such a lexical unit are the gestures, which, as we have already seen, can contrast with one another by means of differing descriptor values. In addition, gestural scores for different lexical items can contrast in terms of the presence vs. absence of particular gestures (Browman & Goldstein, 1986, in press; Goldstein & Browman, 1986). Notice that one implication of taking gestures as basic units is that the resulting lexical representation is inherently underspecified, that is, it contains no specifications for irrelevant features. When a given articulator is not involved in any specified gesture, it is attracted to a 'neutral' position specific to that articulator (Saltzman et al., 1988; Saltzman & Munhall, 1989).

The fact that each gesture has an extent in time, and therefore can overlap with other gestures, has a variety of phonological and phonetic consequences. Overlap between invariantly specified gestures can automatically generate contextual variation of superficially different sorts: (1) acoustic noninvariance, such as the different formant transitions that result when an invariant consonant gesture overlaps different vowel gestures (Liberman & Mattingly, 1985); (2) allophonic variation, such as the nasalized vowel that is produced by overlap between a syllable-final velic opening gesture and the vowel gesture (Krakow, 1989); and (3) various kinds of 'coarticulation,' such as the context-dependent vocal tract shapes for reduced schwa vowels that result from overlap by the neighboring full vowels (Browman & Goldstein, 1989; Fowler, 1981). Here, however, we will focus on the implications of directly representing overlap among phonological units for the phonological/phonetic alternations in fluent speech.

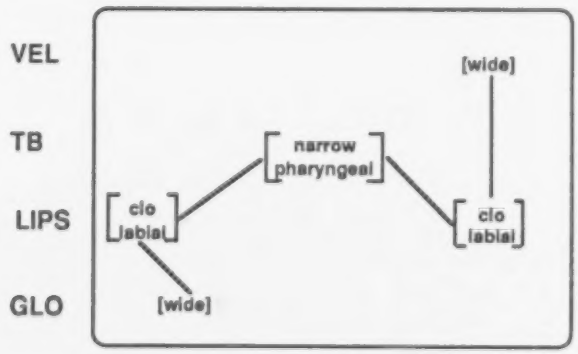


Figure 4. Gestural score for *palm* [pɑm] using point notation, with association lines added.

Browman and Goldstein (1987) have proposed that the wide variety of differences that have been observed between the canonical pronunciation of a word and its pronunciation in fluent contexts (e.g., Brown, 1977; Shockey, 1974) all result from two simple kinds of changes to the gestural score: (1) reduction in the magnitude of individual gestures (in both time and space) and (2) increase in overlap among gestures. That paper showed how these two very general processes might account for variations that have traditionally been described as segment deletions, insertions, assimilations and weakenings. The reason that increased overlap, in particular, can account for such different types of alternations is that the articulatory and acoustic consequences of increased overlap will vary depending on the nature of the overlapping gestures. We will illustrate this using some of the examples of deletion and assimilation presented in Browman and Goldstein (1987), and then compare the gestural account with the treatment of such examples in non-linear phonological theories that do not directly represent overlap among phonological units.

Let us examine what happens when a gestural score is varied by increasing the overlap between two oral constriction gestures, for example from no overlap to complete synchrony. This 'sliding' will produce different consequences in the articulatory and acoustic output of the model, depending on whether the gestures are on the same or different articulatory tiers, i.e., whether they employ the same or different tract variables. If the gestures are on different articulatory tiers, as in the case of a LIPS closure and a tongue tip (TT) closure, then the resulting tract variable motion for each gesture will be unaffected by the other concurrent gesture. Their tract variable goals will be met, regardless of the amount of overlap. However, with sufficient overlap, one gesture may completely obscure the other acoustically, rendering it inaudible. We refer to this as gestural 'hiding.' In contrast, when two gestures are on the same articulatory tier, for example, the tongue tip constriction gestures associated with /t/ and /θ/, they cannot overlap without perturbing each others' tract variable motions. The two gestures are in competition—they are attempting to do different tasks with the identical articulatory structures. In this case, the dynamical parameters for the two overlapping gestural control regimes are 'blended' (Saltzman et al., 1988; Saltzman & Munhall, in press).

Browman and Goldstein (1987) presented examples of articulations in fluent speech that showed the hiding and blending behavior predicted by the model. Examples of hiding are transcribed in (5).

(5) (a) /pəˠfəkt ˈmɛməˠri/ → [pəˠfəkˈmɛməˠri]
(b) /sɛvn̩ ˈplʌs/ → [sɛvm̩ ˈplʌs]

In (5a), careful listening to a speaker's production of *perfect memory* produced in a fluent sentence context failed to reveal any audible /t/ at the end of *perfect,* although the /t/ was audible when the two words were produced as separate phrases in a word list. This deletion of the final /t/ is an example of a general (variable) process in English that deletes final /t/ or /d/ in clusters, particularly before initial obstruents (Guy, 1980). However, the articulatory data for the speaker examined in Browman and Goldstein, (1987) showed that nothing was actually deleted from a gestural viewpoint. The alveolar closure gesture at the end of 'perfect' was produced in the fluent context, with much the same magnitude as when the two words were produced in isolation, but it was completely overlapped by the constrictions of the preceding velar closure and the following labial closure. Thus, the alveolar closure gesture was acoustically hidden. This increase in overlap is represented in Figure 5, which shows the (partial) gestural scores posited for the two versions of this utterance, based on the observed articulatory movements. The gestures for the first syllable of *memory* (shown as shaded boxes) are well separated from the gestures for the last syllable of *perfect* (shown as unshaded boxes) in the word list version in Figure 5a, but they slide earlier in time as shown in Figure 5b, producing substantial overlap among three closure gestures. Note that these three overlapping gestures are all on separate tiers.

In (5b), *seven plus* shows an apparent assimilation, rather than deletion, and was produced when the phrase was produced at a fast rate. Assimilation of final alveolar stops and nasals to a following labial (or velar) stop is a common connected speech process in English (Brown, 1977; Gimson, 1962). Here again, however, the articulatory data in Browman and Goldstein (1987) showed that the actual change was 'hiding' due to increased overlap: the alveolar closure gesture at the end of *seven* was still produced by the speaker in the 'assimilated' version, but it was hidden by the preceding labial fricative and following labial stop.

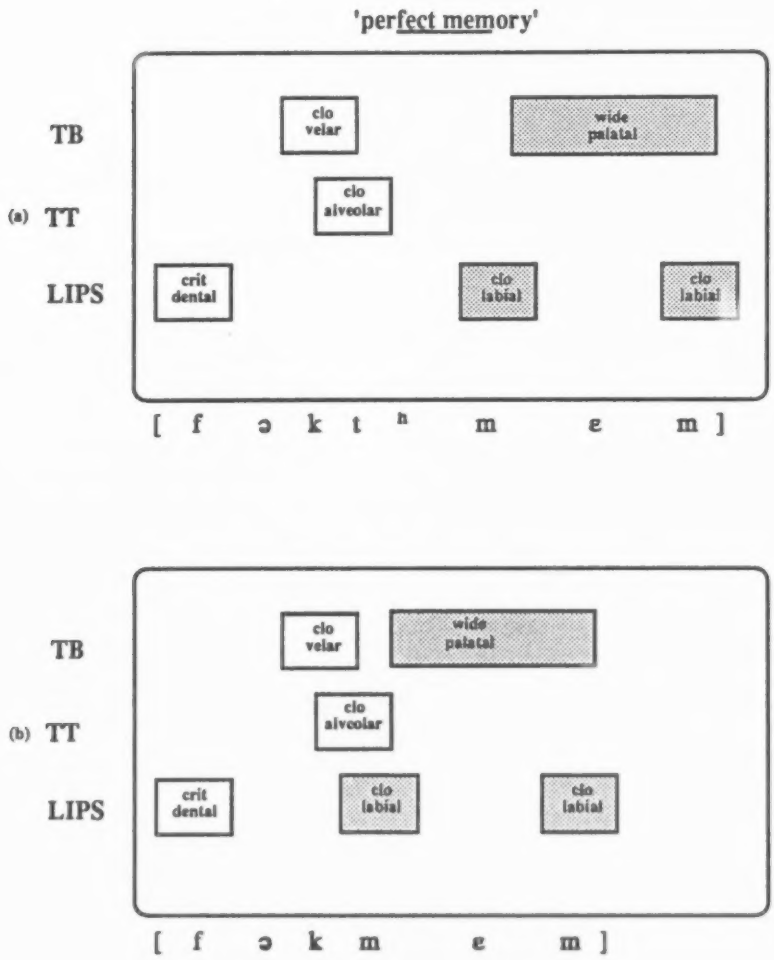


Figure 5. **Partial gestural score for two versions of *perfect memory*, posited from observed articulator movements. Last syllable of *perfect* shown in unshaded boxes; first syllable of *memory* shown in shaded boxes. Only oral tiers are shown. *(a)* Words spoken in word list; *(b)* words spoken as part of fluent phrase.**

The changes in the posited gestural score can be seen in Figure 6. Because the velum lowering gesture (VEL [wide]) at the end of *seven* in Figure 6b overlaps the labial closure, the hidingis perceived as assimilation rather than dele tion. Evidence for such hid-den gestures (in some cases having reduced magnitude) has also been provided by a number of electropalato graphic studies, where they have been taken as evidence of 'partial assimilation' (Barry, 1985; Hardcastle & Roach, 1979; Kohler, 1976 (for German)). Thus, from a gestural point of view, deletion (5a) and assimilation (5b) of final alveolars may involve exactly the same process—increase of overlap resulting in a hidden gesture.

When overlap is increased between two gestures on the same articulatory tier, rather than on different articulatory tiers as in the above examples, the increased overlap results in blending between the dynamical parameters of the two gestures rather than hiding. The trajectories of the tract variables shared by the two gestures are affected by the differing amounts of overlap. Evidence for such blending can be seen in examples like (6).

(6) /tɛn θimz/ → [tɛn̪θimz]

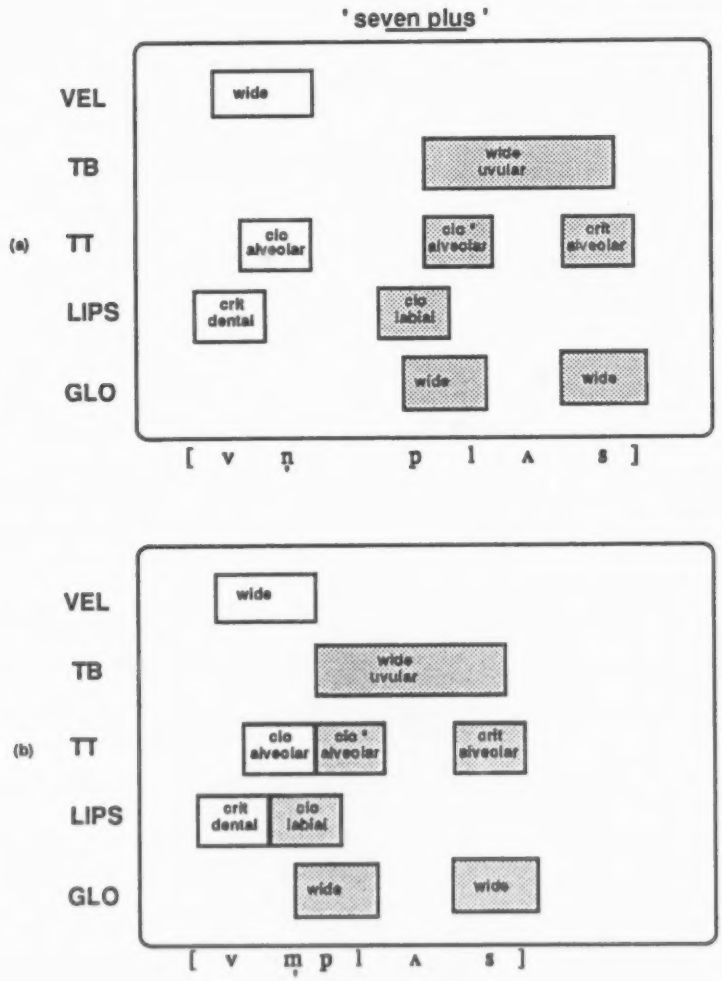


Figure 6. **Partial gestural score for two versions of *seven plus*, posited from observed articulator movements. Last syllable of *seven* shown in unshaded boxes; *plus* shown in shaded boxes. The starred [alveolar clo] gesture indicates that laterality is not represented in these scores. *(a)* Spoken at slow rate; *(b)* spoken at fast rate.**

The apparent assimilation in this case, as well as in many other cases that involve conflicting requirements for the same tract variables, has been characterized by Catford (1977) as involving an accommodation between the two units. This kind of accommodation is exactly what is predicted by parameter blending, assuming that the same underlying mechanism of increased gestural overlap occurs here as in the examples in (5). The (partial) gestural scores hypothesized for (6), showing the increase in overlap, are displayed in Figure 7. The hypothesis of gestural overlap, and consequent blending, makes an specific prediction: the observed motion of the TT tract variables resulting from overlap and blending should differ from the motion exhibited by either of individual gestures alone. In particular, the location of the constriction should not be identical to that of either an alveolar or a dental, but rather should fall somewhere in between. If this prediction is confirmed, then three (superficially) different fluent speech processes—deletion of final alveolar stops in clusters, assimilation of final alveolar stops and nasals to following labials and velars, and assimilation of final alveolar stops to other tongue tip consonants—can all be accounted for as the consequence of increasing overlap between gestures in fluent speech.

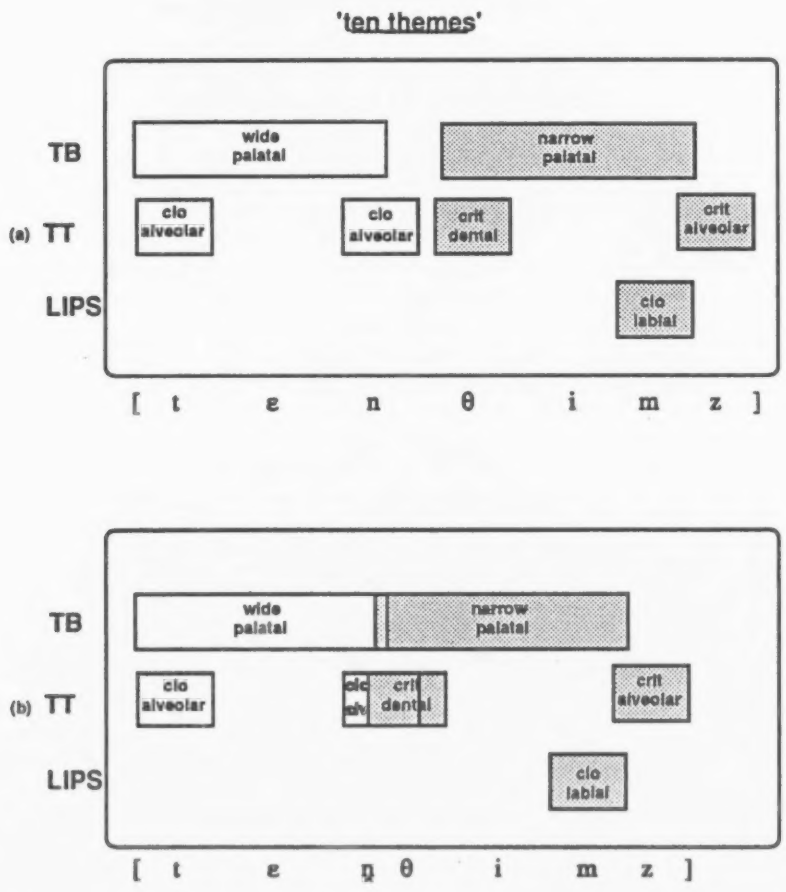


Figure 7. Hypothesized gestured score for two versions of *ten themes*. *ten* shown in unshaded boxes; *themes* shown in shaded boxes. Only oral tiers are shown. *(a)* Spoken with pause between words; *(b)* spoken as part of fluent phrase.

How does the gestural analysis of these fluent speech alternations compare with analyses proposed by other theories of non-linear phonology? Assimilations such as those in (6) have been analyzed by Clements (1985) as resulting from a rule that operates on sequences of alveolar stops or nasals followed by [+coronal] consonants. The rule, whose effect is shown in (7), delinks the place node of the first segment and associates the place node of the second segment to the first (by spreading).

(7) Manner tier [- cont] [+ cons]

Supralaryngeal tier

Place tier [+ cor, - ant] [+ cor]

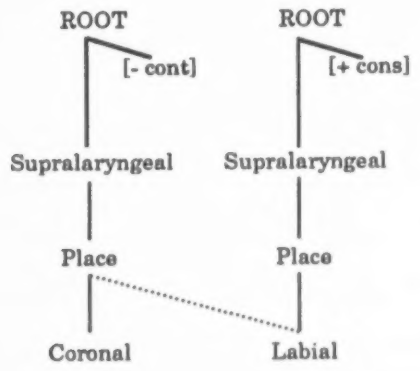
Since the delinked features are assumed to be deleted, by convention, and not realized phonetically, the analysis in (7) predicts that the place of articulation of the assimilated sequence should be indistinguishable from that of the second consonant when produced alone. This claim differs from that made by the blending analysis, which predicts that the assimilated sequence should show the influence of both consonants. These conflicting predictions can be directly tested.

The overlap analysis accounts for a wider range of phenomena than just the assimilations in (6). The deletion of final alveolars (in clusters) and the assimilation of final alveolars to following stops in (5) were also shown to be cases of hiding due to increased gestural overlap. Clement's analysis does not handle these additional cases, and cannot be extended to do so without major reinterpretation of autosegmental formalisms. To see

that this is the case, suppose Clements' analysis is extended by eliminating the [+coronal] requirement on the second segment (for fluent speech). This would produce an assimilation for a case like (5b), but it would not be consistent with the data showing that the alveolar closure gesture is, in fact, still produced. To interpret a delinked gesture as one that is articulatorily produced, but auditorily hidden, would require a major change in the assumptions of the framework.

Within the framework of Sagey (1986), the cases of hiding in (5) could be handled, but the analysis would fail to capture the parallelism between these cases and the blending case in (6). In Sagey's framework, there are separate class nodes for each of the independent oral articulators, corresponding to the articulatory tiers. Thus, an assimilation like that in (5b) could be handled as in (8): the labial node of the second segment is spread to the preceding segment's place node, effectively creating a 'complex' segment involving two articulations.

(8)



However, the example in (6) could not be handled in this way. In Sagey's framework, complex segments can only be created in case there are different articulator nodes, whereas in (6), the same articulator is involved. Thus, an analysis in Sagey's framework will not treat the examples in (5) and (6) as resulting from a single underlying process. If the specific prediction made by the overlap and blending analysis for cases like (6) proves correct, this would be evidence that a unitary process (overlap) is indeed involved—a unity that is directly captured in the overlapping gesture approach, but not in Sagey's framework.

Finally we note that, in general, the gestural approach gives a more constrained and explanatory account of the casual speech changes. All changes are hypothesized to result from two simple mechanisms, which are intrinsically related to the talker's goals of speed and fluency—reduce the size of individual gestures, and increase their overlap. The detailed changes that emerge from these processes are epiphenomenal consequences of the 'blind' application of these principles, and they are explicitly generated as such by our model. Moreover, the gestural approach makes predictions about the kinds of fluent speech variation expected in other languages. Given differences between languages in the canonical gestural scores for lexical items (due to language-specific phasing principles), the same casual speech processes are predicted to have different consequences in different languages. For example, in a language such as Georgian in which stops in the canonical form are always released, word-finally and in clusters (Anderson, 1974), the canonical gestural score should show less overlap between neighboring stops than is the case in English. The gestural model predicts that overlap between gestures will increase in casual speech. But in a language such as Georgian, an increase of the same magnitude as in English would not be sufficient to cause hiding. Thus, no casual speech assimilations and deletions would be predicted for such a language, at least when the increase of gestural overlap is the same magnitude as for English.

The usefulness of internal duration and overlap among phonological elements has begun to be recognized by phonologists (Hammond, 1988; Sagey, 1988). For example, Sagey (1986, pp. 20-21) has argued that phonological association lines, which serve to link the features on separate tiers, 'represent the relation of overlap in time... Thus... the elements that they link...must have internal duration.' However, in nongestural phonologies, the consequences of such overlap cannot be evaluated without additional principles specifying how overlapping units interact. In contrast, in a gestural phonology, the nature of the interaction is implicit in the definition of the gestures themselves, as dynamical control regimes for a (model) physical system. Thus, one of the virtues of the gestural approach is that the *consequences* of overlap are tightly constrained and made explicit in terms of a physical model. Explicit predictions (e.g., about the relation between the constrictions formed by the tongue tip in *themes* and in *ten themes*, as well as about language differences) can be made that test hypotheses about phonological structures.

In summary, a gestural phonology characterizes the movements of the vocal tract articulators during speech in terms of a minimal set of discrete gestural units and patterns of spatiotemporal organization among those units, using dynamic and articulatory models that make explicit the articulatory and acoustic consequences of a particular gestural organization (i.e., gestural score). We have argued that a number of phonological properties of utterances are inherent in these explicit gestural constellations, and thus do not require postulation of additional phonological structure. Distinctiveness (Browman & Goldstein, 1986, in press) and syllable structure (Browman & Goldstein, 1988) can both be seen in gestural scores. In addition, a number of postlexical phonological alternations (Browman & Goldstein, 1987) can be better described in terms of gestures and changes in their organization (overlap) than in terms of other kinds of representations.

2 RELATION BETWEEN GESTURAL STRUCTURES AND FEATURE GEOMETRY

In the remainder of this paper, we look more closely at the relation between gestural structures and recent proposals of feature geometry. The comparison shows that there is much overall similarity and compatibility between feature geometry and the geometry of phonological gestures (§ 2.1). Nevertheless, there are some differences. Most importantly, we show that the gesture is a cohesive unit, that is, a coordinated action of a set of articulators, moving to achieve a constriction that is specified with respect to its *degree* as well as its location. We argue that the gesture, and gestural scores, could usefully be incorporated into feature geometry (§ 2.2). The gestural treatment of constriction degree as part of a gestural unit leads, however, to an apparent disparity with how manner features are currently handled in feature geometry. We conclude, therefore (§ 3) by proposing a hierarchical *tube geometry* that resolves this apparent disparity, and that also, we argue, clarifies the nature of manner features and how they should be treated within feature geometry, or any phonological approach.

2.1 Articulatory geometry and gestural descriptors

In this section, the details of the gestural descriptor structures—the distinctive categories of the tract variable parameters—are laid out. Many aspects of these structures are rooted in long tradition. Jespersen (1914), for example, suggested an 'analphabetic' system of specifying articulatory place along the upper tract, the articulatory organ involved, and the degree of constriction. Pike (1943), in a more elaborated analphabetic system, included variables for impressionistic characterization of the articulatory movements, including 'crests,' 'troughs,' and 'glides' (movements between crests and troughs). More recently, various authors (e.g., Campbell, 1974; Halle, 1982; Sagey, 1986; Venneman & Ladefoged, 1973) have argued that phonological patterns are often formed on the basis of the moving articulator used (the articulator set, in the current system). The gestural structures described in this paper differ from these accounts primarily in the explicitness of the functional organization of the articulators into articulatory gestures, and in the use of a dynamic model to characterize the coordinated movements among the articulators.

The articulatory explicitness of the gestural approach leads to a clear-cut distinction between features of input and features of output. That is, a feature such as 'sonority' has very little to do with articulation, and a great deal to do with acoustics (Ladefoged, 1988a,b). This difference can be captured by contrasting the *input* to the speech production mechanism—the individual gestures in the lexical entry—and the *output*—the articulatory, aerodynamic and acoustic consequences of combining several gestures in different parts of the vocal tract. The gestural descriptors, then, characterize the input mechanism—they are the 'features' of a purely articulatory phonology. Most traditional feature systems, however, represent a conflation of articulatory and acoustic properties. In order to avoid confusion with such combined feature systems, and in order to emphasize that gestural descriptors are purely articulatory, we retain the non-standard terminology developed in the computational model for the category names of the gestural descriptor values.

In § 2.1.1, we discuss the descriptors corresponding to the articulator sets and show how they are embedded in a hierarchical articulatory geometry, comparable to recent proposals of feature geometry. In §§ 2.1.2-5, we examine the other descriptors in turn, demonstrating that they can be used to define natural classes, and showing how the differences

between these descriptors and other feature systems stem from their strictly articulatory and/or dynamic status.

2.1.1 Anatomical hierarchy and articulator sets

The gestural descriptors listed in Figure 2 are not hierarchically organized. That is, each descriptor occupies a separate dimension describing the movement of a gesture. However, there is an implicit hierarchy for the sets of articulators involved. This can be seen in Figure 8, which redisplays (most of) the inventory of articulator sets and associated parameter dimensions from Figure 2, and adds a grouping of gestures into a hierarchy based on articulatory independence. Because the gestures characterize movements within the vocal tract, they are effectively organized by the anatomy of the vocal tract. TT and TB gestures share the individual articulators of tongue body and jaw, so they define a class of Tongue gestures. Similarly, both LIPS and Tongue gestures use the jaw, and are combined in the class of Oral gestures. Finally, Oral, velic (VEL), and glottal (GLO) constitute relatively independent subsystems that combine to form a description of the overall Vocal Tract. This anatomical hierarchy constitutes, in effect, an articulatory geometry.

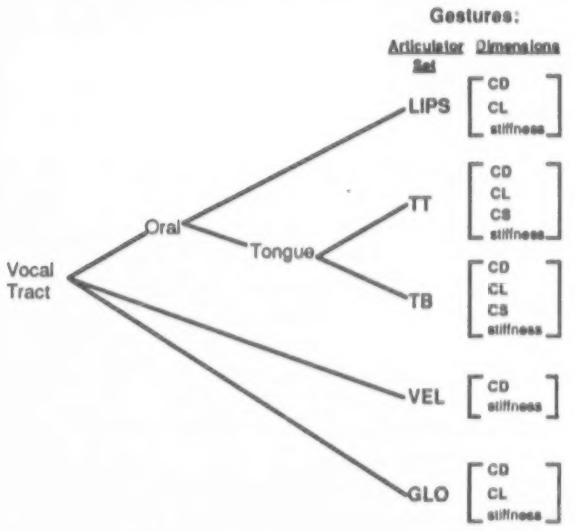


Figure 8. Articulatory geometry tree.

The importance of an articulatory geometry has repeatedly been noted, for example in the use of articulatory subsystems by Abercrombie (1967) and Anderson (1974). More recently, it has formed the basis of a number of proposals of feature geometry (Clements, 1985; Ladefoged, 1988a,b; Ladefoged & Halle, 1988; McCarthy, 1988; Sagey, 1986). The hierarchy of the moving articulators is central to all these proposals of featural organization, although the evidence discussed may be phonological patterns rather than physiological structure. Leaving aside manner features (to be discussed in §§ 2.2.1 and 3), the various hierarchies differ primarily in the inclusion of a Supralaryngeal node in the feature geometries of Clements (1985) and Sagey (1986), and a Tongue node in the articulatory geometry diagrammed in Figure 8.

There is no Supralaryngeal node included in the articulatory geometry of Figure 8, because this geometry effectively organizes the vocal tract *input*. As we will argue in § 3, the Supralaryngeal node is important in characterizing the *output* of the vocal tract. However, using the criterion of articulatory independence, we see no anatomical reason to combine any of the three major subsystems into a higher node in the input hierarchy. Rather than being part of a universal geometry, further combinations of the laryngeal, velic, and Oral subsystems into class nodes such as Supralaryngeal, or Central (Oral) vs. Peripheral (velic and laryngeal), should be invoked where necessitated by language-particular organization. For example, the central-peripheral distinction was argued to exist for Toba Batak (Hayes, 1986).

The Tongue node in Figure 8 is proposed on anatomical grounds, again using the criterion of articulatory independence (and relatedness). That is, the TT articulator set shares two articulators with the TB articulator set—the tongue body and the jaw. Put another way, the tongue is an integral structure that is clearly separate from the lips. There is some evidence for this node in phonological patterns as well. A number of articulations made with the tongue cannot be clearly categorized as being made with either TT or TB. For example, laterals in Kuman (Papua New Guinea) alternate between velars and coronals (Lynch, 1983, cited in McCarthy, 1988). They are always Tongue articulations, but are not categorizable as exclusively TT or TB. Rather, they require some reference to both articulator sets. This is also the case for English laterals, which can alternate between syllable-initial coronals and syllable-final velars (Ladefoged, 1982).

Palatal and palatoalveolar consonants are another type of articulation that falls between TT and TB articulations (Keating, 1988; Ladefoged &

Maddieson, 1986; Recasens, in preparation). Keating (1988) suggests that the intermediate status of palatals—partly TT and partly TB—can be handled by treating palatal consonants as complex segments, represented under both the tongue body and tongue tip/blade nodes. However, without the higher level Tongue node, such a representation equates complex segments such as labio-velars, consisting of articulations of two separate articulators (lips and tongue), with palatals, arguably a single articulation of a single predorsal region of the tongue (Recasens, in preparation). With the inclusion of the Tongue node, labio-velars and palatals would be similar in being double Oral articulations, but different in being LIPS plus Tongue articulations vs. two Tongue articulations. Thus, the closeness of the articulations is reflected in the level of the nodes.

This evidence is suggestive, but not conclusive, on the role of the Tongue node in phonological patterns. We predict that more evidence of phonological patterns based on the anatomical interdependence of the parts of the tongue should exist. One type of evidence should result from the fact that one portion of the tongue cannot move completely independently of the other portions. This lack of independence can be seen, for example, in the suggestion that the tongue body may have a characteristically more backed shape for consonants using the tip/blade in dentals as opposed to alveolars (Ladefoged & Maddieson 1986; Stevens et al., 1986). We would expect to find blocking rules based on such considerations.

2.1.2 Constriction degree (CD)

CD is the analog within the gestural approach of the manner feature(s). However, it is crucial to note that, unlike the manner classes in feature geometry, CD is first and foremost an attribute of the gesture—the constriction made by the moving set of articulators—and therefore, at the gestural level, is solely an articulatory characterization. (This point will be elaborated in § 2.2.1). In our model, CD is a continuum divided into the following discrete ranges: [closed], [critical], [narrow], [mid], and [wide] (Ladefoged, 1988b, refers to such a partitioning of a continuum as an 'ordered set' of values). The two most closed categories correspond approximately to acoustic stops and fricatives; the names used for these categories indicate that these values are articulatory rather than acoustic. Thus, the second degree of constriction, labelled [critical], indicates that critical degree of constriction for a gesture at which some particular aerodynamic consequences could obtain if there were appropriate air flow and muscular tension. That is, the critical constriction value permits frication (turbulence) or voicing, depending on the set of articulators involved (oral or laryngeal). Similarly, [closed] refers to a tight articulatory closure for that particular gesture; the overall state of the vocal tract might cause this closure to be, acoustically, either a stop or a sonorant. This, in turn, will be determined by the combined effects of the concurrently active gestures. § 3 will discuss in greater detail how we account for natural classes that depend on the consequences of combining gestures.

The categorical distinctions among [closed], [critical], and the wider values (as a group) are clearly based on quantal articulatory-acoustic relations (Stevens, 1972). The basis for the distinction among wider values is not as easy to find in articulatory-acoustic relations, although [narrow] might be identified with Catford's (1977) [approximant] category. Catford is also careful to define 'articulatory stricture' in solely articulatory terms. He defines [approximant] as a constriction just wide enough to yield turbulent flow under high airflow conditions such as in open glottis for voicelessness, but laminar flow under low airflow conditions such as in voicing. The other descriptors are required to distinguish among vowels. For example, contrasts among front vowels differing in height are represented as [palatal] constriction locations (cf. Wood, 1982) with [narrow], [mid], or [wide] CDs, where these categories might be established on the basis of sufficient perceptual and articulatory contrast in the vowel system (Lindblom, 1986; Lindblom et al., 1983). If additional differentiation is required, values are combined to indicate intermediate values (e.g., [narrow mid]). In addition, [wide] vs. [narrow] can be used to distinguish the size of glottal aperture—the CD for GLO—associated with aspirated and unaspirated stops, respectively.

2.1.3 Constriction location (CL)

Unlike CD, which differs from its featural analog of manner by being articulatory rather than acoustic, both CL and its featural analog of place are articulatory in definition. However, CL differs from place features in not being hierarchically related to the articulator set. That is, the set of articulators moving to make a constriction, and the location of that constriction, are two independent (albeit highly related) dimensions of a gesture. Thus, we use the label 'Oral' rather than 'place' in the articulatory hierarchy, not only to emphasize the anatomical

nature of the hierarchy, but also to avoid the conflation of moving articulator and location along the tract that 'place' conveys.

The constriction location refers to the location on the upper or back wall of the vocal tract where a gestural constriction occurs, and thus is separate from, but constrained by, the articulator set moving to make the constriction. Figure 9 expands the articulator sets, the CL descriptor values and the possible relations between them: the locations where a given set of articulators can form a constriction. Notice that each articulator set maps onto a subset of the possible constriction locations (rather than all possible CLs), where the subset is determined by anatomical possibility. Notice also that there is a non-unique mapping between articulator sets and CL values. For example, both TT and TB can form a constriction at the hard palate; indeed, it may be possible for TB to form a constriction even further forward in the mouth. Thus constriction location cannot be subsumed under the moving articulator hierarchy, contrary to the proposals by Ladefoged (1988a) and Sagey (1986), among others.

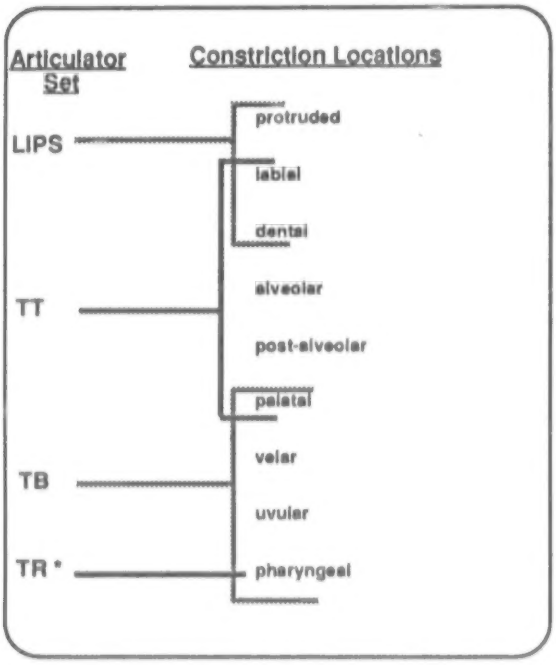


Figure 9. **Possible mappings between articulator sets and constriction locations.**

This can be seen most clearly in the case of the [labial] and [dental] locations, where either the LIPS or TT articulator sets can form constrictions, as shown in (9) (the unusual linguo-labials are discussed in Maddieson 1987). (A similar matrix is presented in Ladefoged & Maddieson, 1986, but not as part of a formal representation.) That is, the [labial] constriction location cannot be associated exclusively with the LIPS articulator set. Similarly, the [dental] constriction location cannot be associated exclusively with the TT articulator set. Thus, we view CL as an independent cross-classifying descriptor dimension whose values cannot be hierarchically subsumed uniquely under particular articulator sets.

(9)

		Articulator set	
		LIPS	TT
CL	labial	bilabial	linguo-labial
	dental	labio-dental	dental

We use multivalued (rather than binary-valued) CL descriptors, following Ladefoged (1988b), since the CL descriptor values correspond to categorical ranges of the continuous dynamic parameter. In the front of the mouth, the basis for the discrete ranges of CL presumably involve the actual differentiated anatomical landmarks, so that parameter values are tuned with respect to them. There may, in addition, be relatively invariant auditory properties associated with the different locations (Stevens, 1989). For the categories that are further back (palatal and beyond), Wood (1982) hypothesizes that the distinct CLs emerge from the alignment of Stevens' quantal considerations with the positioning possibilities allowed by the tongue musculature. To the extent that this set of descriptor values is too limited, it can, again, be extended by combining descriptors, e.g., [palatal velar], or by using 'retracted' and 'advanced' (*Principles of the IPA*, 1949).

2.1.4 Constriction shape (CS)

In some cases, gestures involving the same articulator sets and the same values of CL and CD may differ in the shape of the constriction, as looked at in the frontal, rather than sagittal, plane. For example, constrictions involving TT gestures may differ as to whether they are formed with the actual tip or blade of the tongue, the shape of the constriction being 'wider' in the third dimension if produced with the blade. The importance of this difference has been built into recent feature systems as apical vs. laminal (Ladefoged, 1988a), or [distributed] (Sagey, 1986).

Some method for controlling such differences needs to be built into our computational model. An additional TT tract variable (TTR) that specifies the orientation (angle) of the tongue tip in the sagittal plane with respect to the CL and CD axes is currently being incorporated into the task dynamic model. It is possible that different settings of this variable will be able to produce the required apical/laminal differences, as well as allowing the sublaminal contact involved in extreme retroflex stops (Ladefoged & Maddieson, 1986).

For TB gestures, an additional tract variable is also required to control cross-sectional shape. One of the relevant shape differences involves the production of laterals, in which at least one of the sides of the tongue does not make firm contact with the molars. Ladefoged (1980) suggests that the articulatory maneuver involved is a narrowing of the tongue volume, so that it is pulled away from the sides of the mouth. Such narrowing is an attractive option for an additional TB shaping tract variable. In this proposal, an alveolar lateral would essentially involve two gestures: a TT closure, and a TB gesture with a [narrowed] value for CS and perhaps defaults for CL and CD. (Some further specification of TBCD would clearly be required for lateral fricatives, as opposed to lateral approximants). In this sense, laterals would be complex constellations of gestures, as suggested by Ladefoged and Maddieson (1986), and similar to the proposal of Keating (1988) for treating palatals as complex segments. Another role for a TB shaping tract variable might involve bunching for rhotics (Ladefoged, 1980). Finally, it is unclear whether an additional shape parameter is required for LIPS gestures. It might be required to describe differences between the two kinds of rounding observed in Swedish (e.g., Lindau, 1978; Linker, 1982). On the other hand, given proper constraints on lip shape (Abry & Böe, 1986), it may be possible to produce all the required lip shapes with just LA and LP.

2.1.5 Dynamic descriptors

The *stiffness (k)* of a gesture is a dynamical parameter, inferred from articulatory motions, that has been shown to vary as a function of gestural CD, stress and speaking rate (Browman & Goldstein, 1985; Kelso et al., 1985; Ostry & Munhall, 1985). In addition, however, we hypothesize that stiffness may be tuned independently, so that it can serve as the primary distinction between two gestures. /j/ and /w/ are two cases in which gestural stiffness as an additional independent parameter may be specified. /j/ is a TB [narrow palatal] gesture, and /w/ is a complex formed by a TB [narrow velar] gesture and a LIPS [narrow protruded] gesture. Our current hypothesis is that these gestures have the same CD and CL as for the corresponding vowels (/i/ and /u/), but that they differ in having an [increased] value of stiffness. This is similar to the articulatory description of glides in Catford (1977). Finally, it is possible that the stiffness value that governs the rate of movement into a constriction is related to the actual biomechanical stiffness of the tissues involved. If so, it would be relevant to those gestures that, in fact, require a specific muscular stiffness: trills and taps likely involve characteristic values of oral gesture stiffness, and pitch control and certain phonation types require specified vocal fold stiffnesses (Halle & Stevens, 1971; Ladefoged, 1988a).

2.2 The representation of phonological units

In § 2.1, we laid out the details of gestural descriptors and how they are organized by an articulatory geometry. Setting aside for the moment the critical aspects of gestures discussed in § 1.3 (internal duration and overlap), the differences between feature geometry and the organization of gestural descriptors have so far been comparatively minor—primarily the proposed Tongue node and the non-hierarchical relation between constriction location and the moving articulators. In this section, we consider two ways in which a gestural analysis suggests a different organization of phonological structure from that proposed by the feature geometries referred to in § 2.1.1. First, we present evidence that gestures function as cohesive units of phonological structure (§ 2.2.1). Second, we discuss the advantages of the gestural score as a phonological notation (§ 2.2.2).

2.2.1 The gesture as a unit in phonological patterns

In Figure 8, gestures are in effect the terminal nodes of a feature tree. Notice that the constriction degree, as well as constriction location, constriction shape and stiffness, is part of the descriptor bundle. That is, constriction degree is specified directly at the articulator node—it is one dimension of gestural movement. Looked at in terms of the gestural score (e.g., Figure 3), successive units on a given oral tier contain specifications for both constriction location and

degree of that articulator set. This positioning of CD differs from that proposed in current feature geometries. While CL and CS (or their analogs) are typically considered to be dependents of the articulator nodes, CD (or its analogs) is not. Rather, some of the closest analogs to CD—[stricture], [continuant], and [sonorant]—are usually associated with higher levels, either the Supralaryngeal node (Clements, 1985) or the Root node (Ladefoged, 1988a,b; Ladefoged & Halle, 1988; McCarthy, 1988; Sagey, 1986). Is there any evidence, then, that the unit of the gesture plays a role in phonological organization? Within the approach of feature geometry, such evidence would consist, for example, of rules in which an articulator set and the degree and location of the constriction it forms either spread or delete together, as a unitary whole.

It is in fact generally assumed, implicitly although not explicitly, that velic and glottal features have this type of unitary gestural organization. That is, features such as [+nasal] and [-voice] combine the constriction degree (wide) along with the articulator set (velum or glottis). Assuming a default specification for these features ([-nasal] and [+voice]), then denasalization consists of the deletion of a nasal gesture, and intervocalic voicing of the deletion of a laryngeal gesture. Additional implicit use of a gestural unit can be found in Sagey's (1986) proposal that [round] be subordinate to the Labial node. This is exactly the organization that a gestural analysis suggests, since [round] is effectively a specification of the degree and nature of the constriction of the lips.

It is in the case of primary oral gestures that proposals positioning CD at the gestural level and those positioning it at higher levels contrast most sharply. The inherent connection among all the component aspects of making a constriction, or in other words, the unitary nature of the gesture, can be seen mostly clearly when oral gestures are deleted, a phenomenon sometimes described as delinking of the Place (or Supralaryngeal) node—debuccalization. In a gestural analysis, when the movement of an articulator to make a constriction is deleted, everything about that constriction is deleted, including the constriction degree.

For example, Thráinsson (1978, cited in Clements & Keyser, 1983) demonstrates that in Icelandic, the productive phenomenon whereby the first of a sequence of two identical voiceless aspirated stops is replaced by [h], consists of deleting the first set of supralaryngeal features. This set of supralaryngeal features corresponds to the unit of an oral gesture; it includes the constriction degree, whether described as [clo], [stop] or [-continuant]. Thus, in this example, the entire oral gesture is deleted.

Another example is cited in McCarthy (in press), using data from Straight (1976) and Lombardi (1987). In homorganic stop clusters and affricates (with an intervening word boundary) in Yucatec Maya, the oral portion of the initial stop is deleted—for example, /k#k/ → /h#k/, and (affricate) /ts#t/ → /s#t/. Again in this case the constriction degree is deleted along with the place, supporting a gestural analysis: the entire oral closure gesture is deleted. Note that McCarthy effectively supports this analysis, in spite of his positing [continuant] as dependent on the Root node, when he says 'a stop becomes a segment with no value for [continuant], which is incompatible with supraglottal articulation.'

Thus, there is some clear evidence in phonological patterning for the association of CD with the articulator node. This aspect of gestural organization is totally compatible with the basic approach of feature geometry, requiring only that CD be linked to the articulator node rather than to a higher level node. In § 3, we will show how this gestural affiliation of CD is also compatible with phonological examples in which CD is seemingly separable from the gesture—where it is 'left behind' when a particular articulator-CL combination is deleted, or where it appears not to spread along with the articulatory set and CL. For now, we turn to a second aspect of the gestural approach that could usefully be incorporated into feature geometry, this one involving the use of the gestural score as a two-dimensional projection of the inherently three-dimensional phonological representation.

2.2.2 Gestural scores, articulatory geometry, and phonological units

The gestural score, in particular the 'point' form in Figure 4, is topologically similar to a nonlinear phonological representation. When combined with the hierarchical articulatory geometry, the gestural score captures several relevant dimensions of a nonlinear representation simultaneously, in a clear and revealing fashion. Figure 10 shows a gestural score for *palm*, using point notation and association lines, combined with the articulatory geometry on the left. Note that the geometry tree is represented on its side, rather than the more usual up-down orientation.

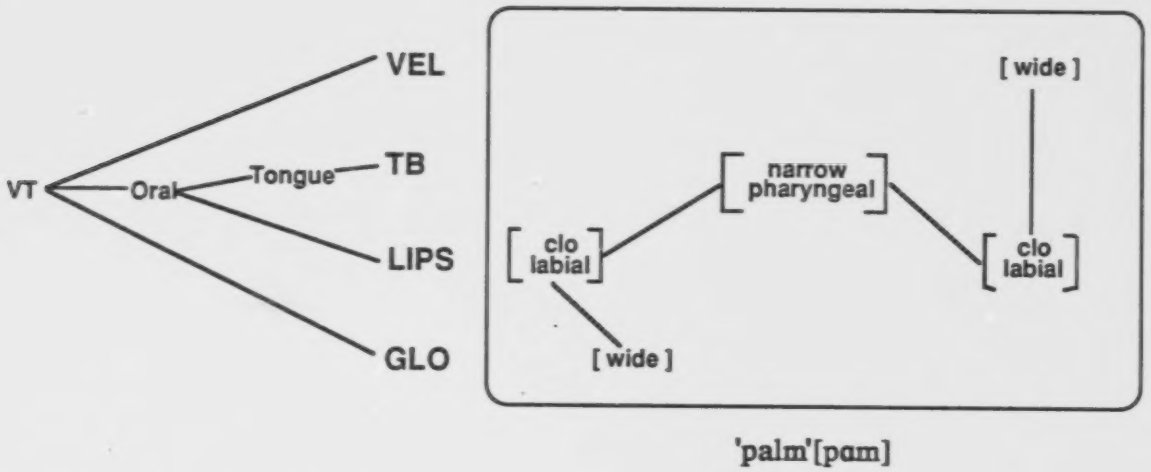


Figure 10. **Gestural score for *palm* [pɑm] in point notation, with articulatory geometry.**

This seemingly trivial point in fact is a very useful consequence of using gestural scores as phonological notation, as we shall see.It permits spatial organization such as the articulatory geometry, which is represented on the vertical axis, to be separated from temporal information including sequencing of phonological units, which is represented on the horizontal axis. Such a separation is particularly useful in those instances in which the gestures in a phonological unit are not simultaneous.

For example, prenasalized stops are single phonological units consisting of a sequence of nasal specifications. Figure 11a depicts a gestural score for prenasalized [mb], which contains a closure gesture on the LIPS tier associated with a sequence of two gestures on the VEL tier, one with CD = [wide] followed by one with CD = [clo]. Double articulations also constitute a single phonological unit. Figure 11b displays a gestural score for [g͡b], which consists of a [clo labial] gesture on the LIPS tier associated with a [clo velar] gesture on the TB tier. Sagey (1986) terms the first type of unit a contour segment, and the second type a complex segment, a terminology that is an apt description of the two figures.

Compare the representations in Figures 11a and 11b to those in Figures 11c and 11d, which are Sagey's (1986) feature geometry specifications of the same two phonological units. In the feature geometry representation, the two types of segments appear to be equally sequential, or equally non-sequential. This is a consequence—an unfortunate consequence—of conflating two uses of branching notation, that of indicating (order-free) hierachical information and that of indicating sequencing information. Thus, in Figure 11d (on the right), the branching lines represent order-free branching in a hierarchical tree. In Figure 11c, however, the branching lines represent the associations between two ordered elements and a single node on another tier. This conflation results from the particular choice of how to project an inherently three-dimensional phonological structure (feature hierarchy × phonological unit constituency × time) onto two dimensions.

In the gestural score, the two-dimensional projection avoids this conflation of the two uses of branching notation. Here nodes always represent gestures and lines are always association (or phasing) lines. Thus any branching, as in Figure 11a, always indicates temporal information about sequencing. The hierarchical information about the articulatory geometry is present in the organization of the articulatory tiers, and is thus represented (once) by a sideways tree all the way at the left of the gestural score (as in Figure 10). In this way, the gestural score retains the virtues of earlier forms of phonological notation in which sequencing information was clearly distinguishable, while also providing the benefits of the spatial geometry that organizes the tiers hierarchically. It would, of course, be possible to adopt this kind of representation within feature geometry.

Finally, the reader may have observed that, in discussions to now, constituency in phonological units has been indicated solely by using association lines between gestures. In particular, there has been no separate representation of prosodic phonological structure. This is not an inherent aspect of a phonology of articulatory gestures; rather, it is the result of our current research strategy, which is to see how much structure inheres directly in the relations among gestures, without recourse to higher level nodes. However, once again, it is possible to integrate the gestural score with other types of phonological representation. To exemplify how the gestural score can be integrated with explicit phonological structure, Figure 12 displays a mapping between the gestural score for *palm* and a simplified version of the prosodic structure of Selkirk (1988), with the articulatory geometry indicated on the left.

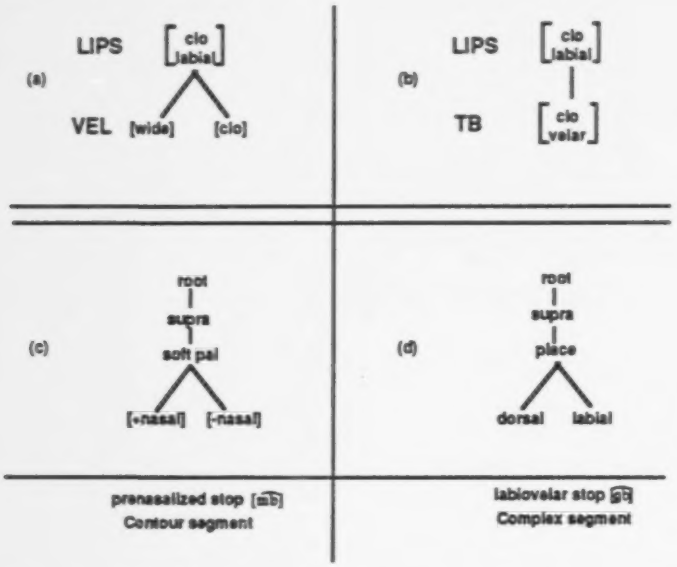


Figure 11. **Comparison of gestural score and feature geometry representations for contour and complex segments.** *(a)* **Gestural score for contour segment [m͡b]; *(b)* gestural score for complex segment [g͡b]; *(c)* feature geometry for contour segment [m͡b]; *(d)* feature geometry for complex segment [g͡b].**

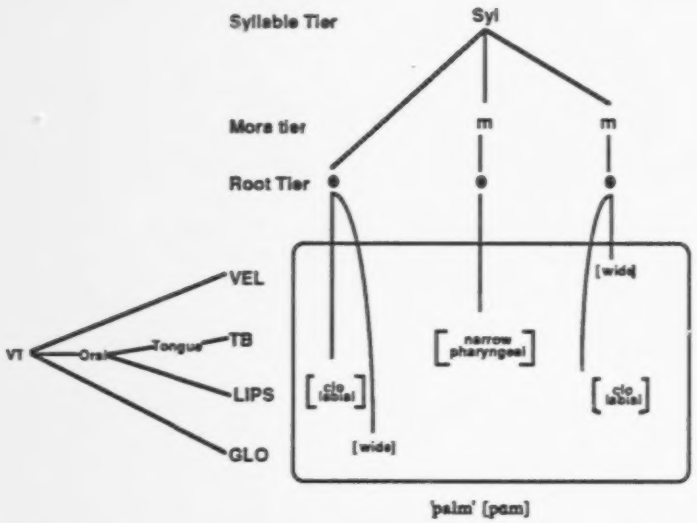


Figure 12. **Gestural score for *palm* [pɑm] in point notation, mapped on to prosodic structure, with articulatory hierarchy on left.**

In short, although gestures and gestural scores originated in a description of articulator movement patterns, they nevertheless provide constructs that are useful for phonological representation. Gestures constitute an organization (combining a moving articulator set with the degree and location of the constriction formed) that can act as a phonological unit. Gestural scores provide a useful notation for nonlinear phonological representations, one that permits phonological constituency to be expressed in the same representation with 'phonetic' order information, which is indicated by the relative positions of the gestures in the temporal matrix. The gestures can be grouped into higher-level units, using either association lines or a mapping onto prosodic structure, regardless of the simultaneity, sequentiality, or partial overlap of the gestures. In addition, the articulatory geometry organizing the tiers can be expressed in the same representation.

3 TUBE GEOMETRY AND CONSTRICTION DEGREE HIERARCHY

So far, we have been treating individual gestures as the terminal nodes of an anatomical hierarchy. From this perspective, articulatory geometry serves as a way of creating natural classes of gestures on anatomical grounds. In this section, we turn to the vocal tract as a whole, and consider how additional natural classes emerge from the combined effect (articulatory, aerodynamic and acoustic consequences) of a set of concurrent gestures. We argue that these natural classes can best be characterized using a hierarchy for constriction degree within the vocal tract based on *tube geometry*.

Rather than viewing the vocal tract as a set of articulators, organized by the anatomy, tube geometry views it as a set of tubes, connected either in series or in parallel. The sets of articulators move within these tubes, creating constrictions. In other words, articulatory gestures occur within the individual tubes. More than one gesture may be simultaneously active; together, these gestures interact to determine the overall aerodynamic and acoustic output of the entire linked set of tubes. Similarities in the tube consequences of a set of different gestures may lead to similar phonological behavior: this is the source of acoustic features (Ladefoged, 1988a) such as [grave] and [flat] (Jakobson, Fant, & Halle, 1969) that organize different gestures (in our terms) according to standing wave nodes in the oral tube (Ohala, 1985; Ohala & Lorentz, 1977).

Within this tube perspective, there is a vocal tract hierarchy that characterizes an instantaneous time slice of the output of the vocal tract. Such a hierarchy may appear identical to the feature geometry of Sagey (1986). There are, however, two crucial differences. Unlike Sagey's root node, which 'corresponds neither to anatomy of the vocal tract nor to acoustic properties' (1986, p. 16), the highest node in the vocal tract hierarchy characterizes the physical state of the vocal tract at a single instant in time. In addition, tube geometry characterizes manner—CD—at more than just the highest level node. CD is characterized at each level of the hierarchy. Thus, each of the tubes and sets of compound tubes in the hierarchy will have its own effective constriction degree that is completely predictable from the CD of its constituents, and hence ultimately from the CD of the currently active gestures. At the vocal tract level, the effective CD of the supralaryngeal tract, taken together with the initiator power provided by the lungs, determines the nature of the actual airflow through the vocal tract: none (complete occlusion), turbulent flow, or laminar flow. This vocal tract hierarchy thus serves as the basis for a hierarchy for CD.

The important point here for phonology is that, in the output system, CD exists simultaneously at all the nodes in the vocal tract hierarchy—it is not isolable to any single node, but rather forms its own CD hierarchy. As we argue in § 3.2, all levels of the hierarchy are potentially important in accounting for phonological regularities, and may form the basis of a natural class. First, however, in § 3.1, we expand on the nature of tube geometry.

3.1 How tube geometry works

Figure 13 provides a pictorial representation of the tubes that constitute the vocal tract, embedded in the space defined by the anatomical hierarchy of articulatory geometry, in the vertical dimension, and by the constriction degree hierarchy of tube geometry, in the horizontal dimension. The constriction action of individual gestures is schematically represented by the small grey disks. Thus, the two dimensions in the figure serve to organize the gestures occurring in the vocal tract tubes, vertically in terms of articulatory geometry, and horizontally in terms of tube geometry. Looking at the tube structure in the center, we can see that the vocal tract

branches into three distinct tubes or airflow channels: a Nasal tube, a Central tongue channel and a Lateral tongue channel. The complex is terminated by GLO gestures at one end. At the other end, the Central and Lateral channels are together terminated by LIPS gestures.

The tube geometry tree, shown at the top of the figure, reflects the combinations of the tubes and their terminators. The tube level of the tree has five nodes, one for each of the three basic tubes and the two terminators. Within each of these basic tubes, the CD will be determined by the CD of the gestures acting within that tube, which are shown as subordinates to the tube nodes. For example, the CD of TT gestures will contribute to determining the CD of the Central tube. TB gestures will contribute to the Central and/or Lateral tube, depending on the constriction shape of the TB gesture. Each of the superordinate nodes corresponds to a tube junction, forming a compound tube from simpler tubes and/or the termination of a tube. Thus, the Central and Lateral tubes together form a compound tube labelled the Tongue tube. The compound Tongue tube is terminated by the LIPS configuration, which combines with the Tongue tube to form an Oral tube. The Oral and Nasal tubes form another compound tube, the Supralaryngeal, which is terminated by GLO gestures to form the overall Vocal Tract compound tube.

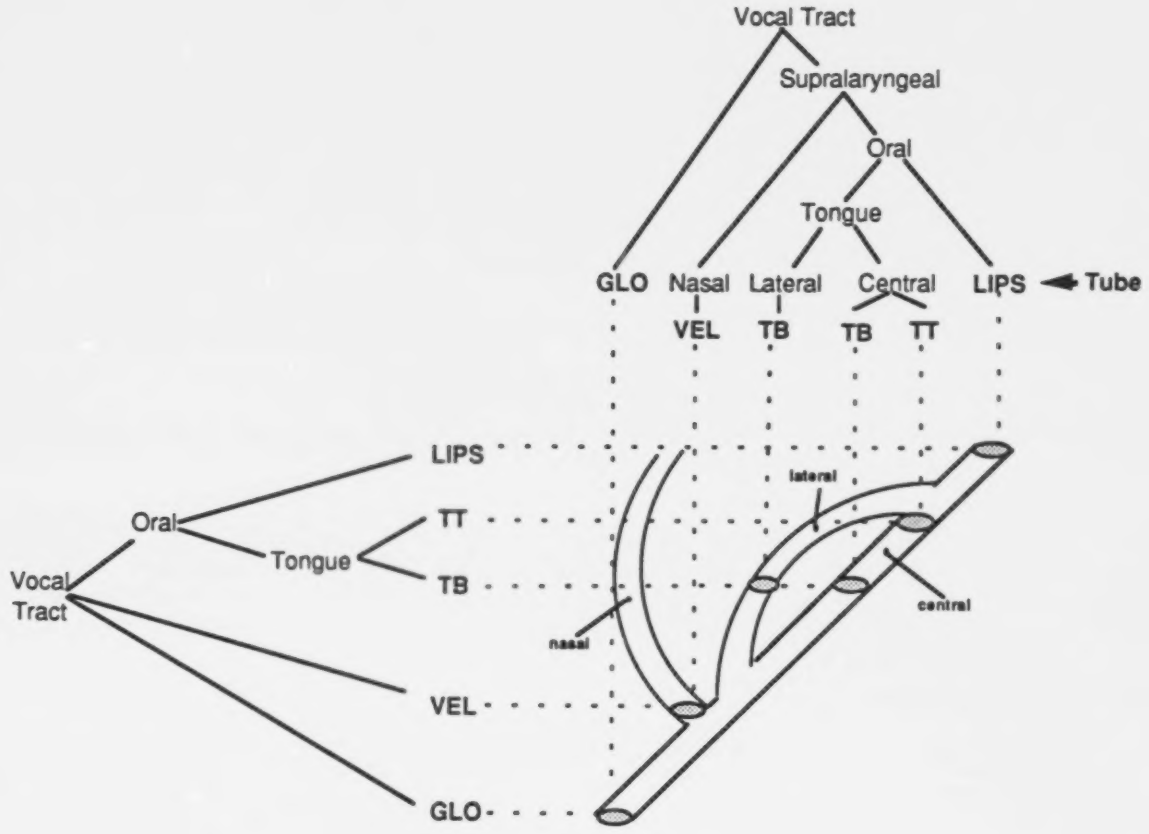


Figure 13. **Vocal tract hierarchy: Articulatory and tube geometry.**

The effective CD at each superordinate node can be predicted from the CD of the tubes being joined and the way they are joined. When tubes are joined in parallel, the effective CD of the compound tube has the CD of the *widest* component tube, that is, the maximum CD.When they are joined in series, the compound tube has the CD of the *narrowest* component tube, that is, the minimum CD. Terminations and multiple constrictions within the same tube work like tubes connected in series. Using these principles, it is possible to 'percolate' CD values from the values for individual gestures up to the various nodes in the hierarchy. Table 1 shows the possible values for CD of individual gestures, and Table 2 shows how to determine the CD values at each successive node up through the Supralaryngeal node, referred to hereafter as the Supra node (the Vocal Tract node will be discussed below).

Table 1. *Possible constriction degree (CD) values at gestural tract variable level : open = = (narrow* OR *mid* OR *wide).*

a. LIPS	[CD]	= clo crit open
b. TT	[CD]	= clo crit open
c. TB	[CD, CS = normal]	= clo crit open
d. TB	[CD, CS = narrowed]	= clo crit open
e. VEL	[CD]	= clo open
f. GLO	[CD]	= clo crit open

GLO [crit] is value appropriate for Voicing

Table 2. *Percolation of CD up through Supralaryngeal node.*

a. Nasal	[CD] = VEL [CD]
Lateral	[CD] = TB [CD, CS = narrowed]
Central	[CD] = MIN (TT [CD], TB [CD, CS = normal])
b. Tongue	[CD] = MAX (Central [CD], Lateral [CD])
c. Oral	[CD] = MIN (Tongue [CD], LIPS [CD])
d. Supra	[CD] = MAX (Oral [CD], Nasal [CD])

The percolation principles follow from aerodynamic considerations. Basically, airflow through a tube system will follow the path of least resistance, as we can illustrate with examples of nasals and laterals. The Oral and Nasal tubes are connected in parallel, forming a compound Supralaryngeal tube. If the Oral tube is [closed] but the Nasal tube is [open], Table 2d indicates that the CD of the combined Supra tube will be the wider of the two openings, i.e., [open], as it is in a nasal stop. To take a less obvious example, consider the case where the Nasal tube is [open], but the Oral tube is [crit], i.e., appropriate for turbulence generation under the appropriate airflow conditions. Table 2d predicts that in this case as well the Supra CD is [open], implying that there will be no turbulence generated. This in turn predicts that nasalized fricatives, which consist of exactly the VEL [open] and Oral [crit] gestures, should not exist. That is, given that, at normal airflow rates, the air in this configuration will tend to follow the path of least resistance through the [open] nasal passage, nasalized fricatives would require abnormal degrees of total airflow in order to generate airflow through the [crit] constriction that is sufficient to produce turbulence.

Ohala (1975) has proposed that such nasalized fricatives are, indeed, rare, precisely because they are hard to produce. As Ladefoged and Maddieson (1986) argue, this could account for alternations in which the nasalized counterparts of voiced fricatives are voiced approximants, such as in Guaraní (Gregores & Suárez, 1967, in Ladefoged & Maddieson, 1986). However, they also present evidence from Schadeberg (1982) for a nasalized labio-dental fricative in Umbundu. This suggests that the percolation principles may be overridden in certain special cases, perhaps by increased airflow settings.

At the highest level, that of the Vocal Tract, the glottal (GLO) CD and stiffness and the Supra CD combine with initiator (pulmonic) action to determine the actual aerodynamic and acoustic characteristics of airflow through the vocal tract. At this point, then, it becomes more appropriate to label the states of the Vocal Tract in terms of the characteristics of this 'output' airflow, rather than in terms of the CD, which is one of its determining parameters. A gross, but linguistically relevant, characterization of the output distinguishes the three states defined in Table 3a: occlusion, noise and resonance. Assuming some 'average' value for initiator power, Table 3b shows how GLO and Supra CD jointly determine these properties. A complete closure of either system results in occlusion. If GLO is [crit] (appropriate position for voicing, assuming also appropriate stiffness), then it combines with an open Supralaryngeal tract to produce resonance. Any other condition produces

noise. In some cases, e.g., GLO [open] and Supra [open], this is weak noise generated at the glottis (aspiration). In other cases, e.g., Supra [crit], the noise is generated in the Oral tube (frication). Further details of the output, such as where the turbulence is generated, and whether voicing accompanies occlusion or noise, are beyond our scope here. Ohala (1983) gives many examples of how the principles involved are relevant to aspects of phonological patterning.

Table 3. *Acoustic consequences at Vocal Tract level.*

a. VT outputs		
occlusion:	no airflow through VT; silence or low-amplitude voicing	
noise:	turbulent airflow	
resonance:	laminar airflow with voicing; formant structure	
b. VT [CD]	= occlusion	/ Supra [clo] OR GLO [clo]
	= resonance	/ Supra [open] AND GLO [crit]
	= noise	/ otherwise

Finally, note that the percolation of CD through levels of the tube geometry can be defined at any instant in time, and depends on the actual size of the constriction within each tube at that point in time, regardless of whether some gesture is actively producing the constriction. Thus, the instantaneous CD is the output consequence of the default values for articulators not under active control, the history of the articulator movements, and the constellation of gestures currently active. Table 4 shows the default value we are assuming for each basic tube and terminator. (The default for Lateral is seriously oversimplified, and does not give the right Lateral CD in the case of Central fricatives, although percolation to the Oral level will work correctly).

Table 4. *Default CD values for basic tubes and terminators.*

Nasal	[CD] = clo
Lateral	[CD] = Central [CD]
Central	[CD] = open
LIPS	[CD] = open
GLO	[CD] = crit

Default values are determined by the effect of the model articulators' neutral configuration within a tube

3.2 The constriction degree hierarchy

The importance of tube geometry for phonology resides in the fact that constriction degree is not isolable to any single level of the vocal tract. Rather, constriction degree exists at a number of levels simultaneously, with the CD at each level in this hierarchy defining a potential natural class. In this section, we present some examples in which CD from different levels is phonologically relevant, and argue that the CD hierarchy is important and clarifying for any featural system, as well as for gestural phonology.

We are proposing (1) that there is a universal geometry for CD in which all nodes in the CD hierarchy are simultaneously present, and (2) that different CD nodes are used to represent different natural classes. This approach differs from that adopted in current feature geometries, in which manner features such as [nasal], [sonorant], [continuant] or [consonantal] are usually represented at a single level in the feature hierarchy. Different geometries, however, choose different levels. For example, [nasal] has been variously considered to be a feature dependent on the highest (Root) node (McCarthy, 1988), on the Supralaryngeal node (via a manner node) (Clements, 1985), or on a Soft Palate node (Ladefoged, 1988a,b; Ladefoged & Halle, 1988; Sagey, 1986). It is possible that this variability in the treatment of [nasal], and other features, reflects the natural variation in the CD hierarchy, such that different phonological phenomena are captured using CD at different levels in the hierarchy.

To see how this might work, consider the four hierarchies in Figure 14, which use the tube geometry at the top of Figure 13 to characterize the linked CD structures for a vowel, a lateral, a nasal, and an oral stop. Since tube geometry plays the same role in the representation as the articulatory geometry in Figure 10, the tube hierarchies are rotated 90 degrees just as the anatomical hierarchy was. The circles are a graphic representation of the CD for each node, with the filled circles indicating CD = [clo], the wavy lines indicating CD = [crit], and the open circles indicating CD = [open] (the symbols indicate occlusion, noise, and resonance, respectively, at the Vocal Tract level). These CD hierarchies express the various classificatory (featural) similarities and dissimilarities among the four segments, but they do so at different levels.

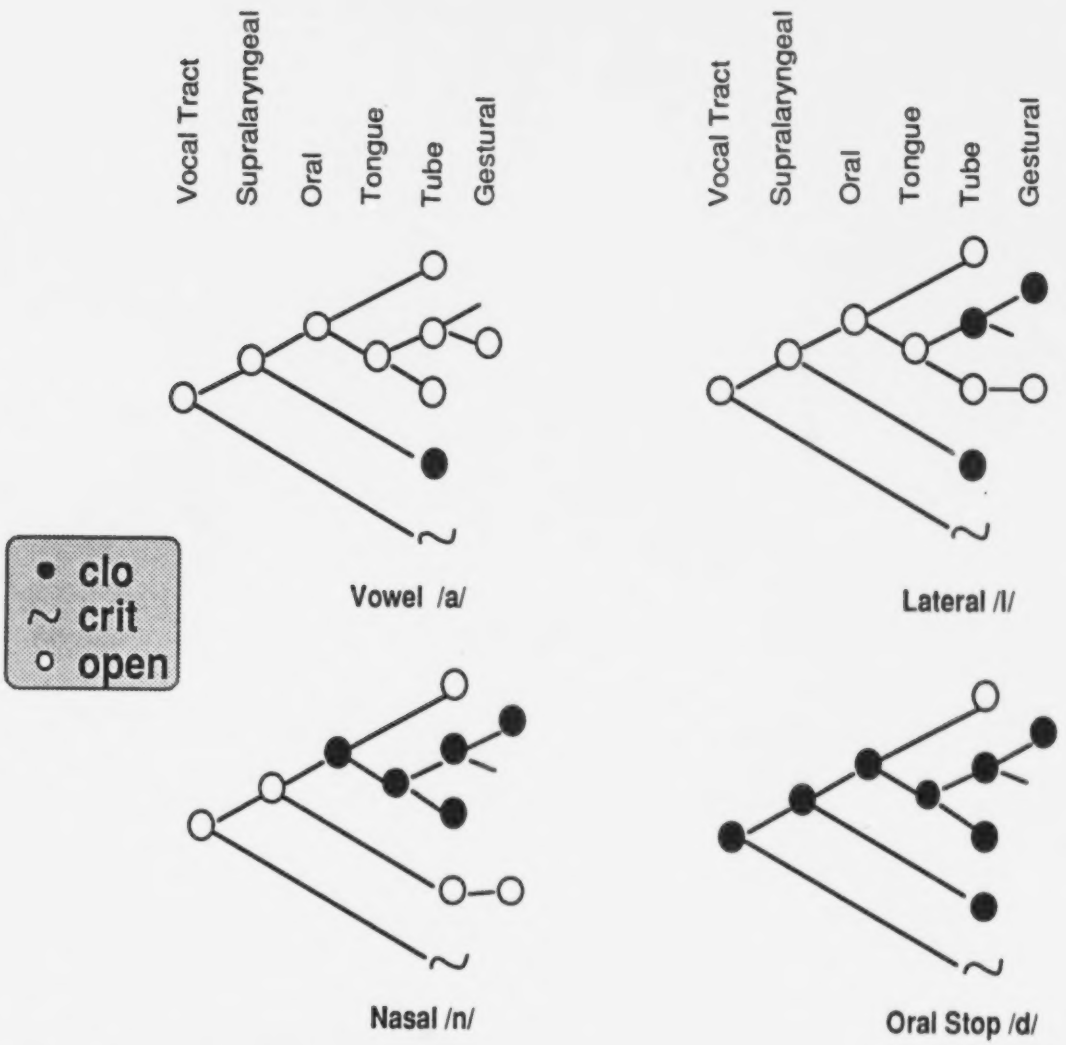


Figure 14. CD hierarchies for vowel, lateral, nasal and oral stop.

For example, the nasal has the same characterization as the vowel and lateral at the Supralaryngeal and Vocal Tract levels, but diverges at lower levels. This aspect of the CD hierarchy thus defines a phonological natural class consisting of nasals, laterals, and vowels, where the class is characterized by Supra [open] and VT ['open' = resonance]. This natural class is typically represented by the feature [sonorant], although different feature systems select one or the other of these levels in their definition of sonority. For example, Ladefoged's (1988a,b) acoustic feature distinction of [sonorant] utilizes the identity of the value in the CD hierarchy at the VT level. That is, for Ladefoged [+sonorant] differs from [-sonorant] in terms of output at the VT level. However, for Chomsky and Halle (1968) and Stevens (1972), it is the identity at the Supralaryngeal level that characterizes sonority, since they consider /h/ and /ʔ/ to be sonorants even though they differ from the nasal, vowel and lateral at the VT level.

At the Oral level (and below), the nasal and stop display the same linked CD structure, which differs from that of the lateral and that of the vowel. That is, nasals and stops form a natural class in having Oral [clo], whereas laterals and vowels form a class in having Oral [open]. This difference in CD resides at a lower level of the hierarchy than the Vocal Tract or Supralaryngeal levels. And at the lowest levels, the Tube and Gestural, the nasal differs from the stop, lateral, and vowel in having both a Nasal [CD] and VEL [CD] that are [open]. Thus, constriction degree at various levels can serve to categorize and distinguish phonological units in different ways.

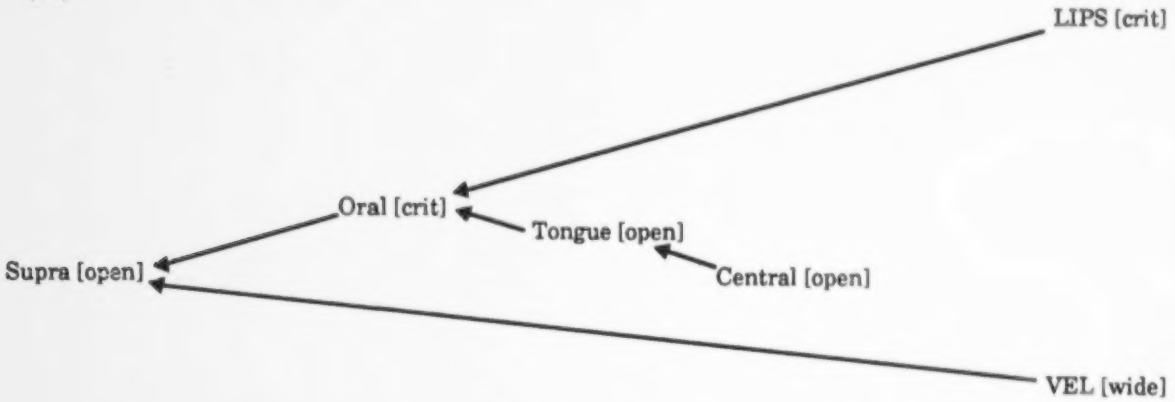
Within traditional feature systems, each of the natural classes described above receives a separate name, which obscures the systematic relation among the various classes. Moreover, even in feature geometry, when manner features are restricted to a single level there is no principled representation of the hierarchical relation among the natural classes. However, feature geometries sometimes incorporate pieces of the CD hierarchy, for example, in the assignment of [nasal] and [continuant] as dependents of two different nodes in the hierarchy (Soft Palate and Root: Sagey, 1986), or the assignment of [sonorant] as part of the Root node itself, but [continuant] as a dependent on the Root node (McCarthy, 1988). Note, however, that hierarchical relations among manner classes have to be stipulated in feature geometries, whereas such relations are inherent in the CD hierarchy. In addition, the percolation principles of tube geometry provide a mechanism for relating values of the different levels to each other, again something that would need to be stipulated in a hierarchy not based on tube geometry.

All the levels of the CD hierarchy appear to be useful for establishing natural classes, and for relating the CD natural classes to one another. In addition, various kinds of phonological patterns, such as phonological alternations, can be examined in light of this hierarchy. In particular, we can investigate whether regularities are best expressed as processes that treat CD as tied to a particular gesture (as an 'input' parameter) or as a consequence of gestural combination at the various 'output' levels of the CD hierarchy. For example, the cases discussed in § 2.2.1 (e.g., /k#k/ → /h#k/) can be best described as processes that delete entire gestures. CD is tied to the gesture and is deleted along with the articulator set and CL. This deletion automatically accounts (by percolation) for the CD changes at other levels of the hierarchy.

In other cases, however, there is much overdetermination—the phonological behavior can be described equally well from more than one perspective. For example, McCarthy (1988) discusses the common instances in which /s/ → /h/, and glottalized consonants /p' t' k'/ → /ʔ/. Both of these examples can be straightforwardly analyzed as deletion of the oral gesture, as in the cases in § 2.2.1. But it is also true that, in both cases, the Vocal Tract output CD is unchanged by the gestural deletion. In the first case, it remains noisy, and in the second it remains an occlusion. Thus, the description of the phonological behavior could also focus on the (apparent) relative independence of the VT [CD] and the articulator set involved—the articulator set(s) changes, but CD does not. The equivalence of the two perspectives results from that fact that the percolated values of VT [CD] are the same for the two gestures alone and for their combination. Thus, deletion of one or the other will not change the VT [CD]. Examples of this kind, of which there are likely to be many, can be viewed from a dual perspective.

One example of process for which it seems (at first glance) that a dual perspective cannot be maintained involves assimilation. Steriade (in press) has argued that assimilation is problematic for the gestural approach, since sometimes only the place, and not the manner, features are assimilated. For example, in Kpelle nasals assimilate in place but not manner to a following stop or fricative, so that /N-f/ becomes a sequence (broadly) transcribed as [mv] (Sagey, 1986). As Steriade points out, this separation of the place and manner features appears to argue against a gestural analysis, in which the assimilation results from increased overlap between the oral gesture for the fricative (LIPS [crit den]) and the velum lowering gesture (VEL [wide]). Since the nasal does not become a nasal fricative, it would appear either that there is no increased overlap, or that the LIPS [crit] gesture changes its CD when it overlaps the velum lowering gesture. From the perspective of the CD hierarchy, the Supra CD of the nasal is the same before and after the assimilation ([open]), and therefore the Supra level (or VT level) is the significant one for CD, rather than the Gestural level. However, the percolation principles from tube geometry predict that the Supra CD will be [open] even if the nasal is overlapped by a fricative gesture, as derived in (10).

(10)

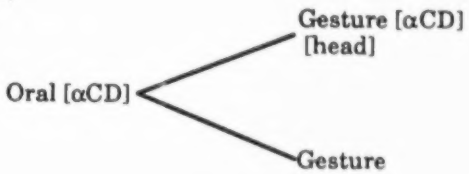


There could also be a TT gesture in the derivation that is either hidden or deleted—the Supra CD would still be [open]. Thus, an increase in gestural overlap could produce both the place assimilation as well as the correct CD at the Supra level, without any change of the Gestural CD. (An articulatory study is needed to determine whether the Gestural CD does indeed change).

The examples presented so far are processes that are either best described with CD attached firmly to a given gesture as it undergoes change (e.g., deletion or sliding), or are at least equally well described that way. However, there are phenomena whose description requires a 'loosening' of the relation between CD and the gesture. One such situation, in which the Gestural CD is effectively separated from its articulator set, occurs in historical change (Browman & Goldstein, in press), in particular in the types of historical change that Ohala (1981) terms 'listener-based' sound changes. Such cases arise particularly when gestural overlap leads to ambiguity in the acoustic signal that the listener can parse into one of two gestural patterns. In such cases, Browman and Goldstein (in press) argue that there is a (historical) 'reassignment' of the constriction degree parameter between overlapping gestures. This analysis was used to account for the /x/ → /f/ changes in English words like *cough*, based on an argument originally put forth by Pagliuca (1982). Briefly, the analysis proposed that the [crit] descriptor for the TB gesture for /x/ was re-assigned to an overlapping LIPS gesture.

A related synchronic situation, involving two overlapping gestures in a complex segment, is discussed in Sagey (1986). Her analysis suggests that manner features must be represented on more than a single node in a feature hierarchy. Specifically, Sagey demonstrates that multiple oral gestures cohering in a complex segment may be restricted to a single distinctive constriction degree. She represents this single distinctive constriction degree at the highest level in the hierarchy, but still must represent which particular articulation bears the contrastive CD. Thus, CD is specified at two different levels. Sagey diagrams the relation between the two levels by a looping arrow drawn between the distinctive level and the 'major' articulator making the distinctive constriction. In the CD hierarchy, the Oral CD would be the lowest possible distinctive level for these double articulations. The gesture that carries the distinctive CD could be marked as [head], and would automatically agree in CD with the mother node, as in (11).

(11)



Moreover, in examples such as this, the percolation principles do not contribute to determining the CD of the distinctive node, except in a negative way to be discussed shortly. Sagey argues that the distinctive CD cannot be predicted from physical principles, since in Margi labio-coronals, the coronal articulation is major, and hence in /p͡s/ the less radical constriction is the distinctive constriction, rather than the more radical one. Thus, the relation between the Oral and Gestural levels in (11) must be a statement about a functional phonological unit, a gestural constellation, rather than a statement

characterizing an instantaneous time slice in terms of tube geometry. That is, only one of the gestures in the constellation may bear a distinctive CD. Nevertheless, the importance of physical overlap relations can also be seen in this example. Maddieson and Ladefoged (1989) have shown that complex segments such as [g͡b] are not completely synchronous, apparently lending support to the distinction between phonetic ordering, on the one hand, and phonological unordering, on the other hand. However, we argue that the ordering of gestures within a complex segment *is* phonologically important in exactly the case of a phonological unit like Margi /p͡s/, where the distinctive value of CD at higher levels is not predicted from the lower level CDs by the percolation principles. If two gestures overlap completely (i.e., are precisely coextensive), and there are no additional phonetic cues of the sort discussed in Maddieson & Ladefoged, then the percolation principles will determine the higher level constriction degree throughout the entire time-course of the phonological unit, and the distinctive CD will fail to be conveyed. For example, in the case of /p͡s/, if the two gestures were precisely aligned, the (distinctive) frication would never appear at the VT level. Only if the gestures are slightly offset can the distinctive CD be communicated.

In general, the CD hierarchy affords a structure within which a typology of phonological processes can be developed, based on the CD level that seems most relevant. It is an interesting research challenge to develop this typology, and to ask how it is related to other ways of categorizing the processes. For example, are there systematic differences in relevant CD level that correlate with whether the process involves spreading (sliding) rules or deletion rules, or with whether the process is prelexical or postlexical?

4 SUMMARY

We have argued that dynamically-defined articulatory gestures are the appropriate units to serve as the atoms of phonological representation. Gestures are a natural unit, not only because they involve task-oriented movements of the articulators, but because they arguably emerge as prelinguistic discrete units of action in infants. The use of gestures, rather than constellations of gestures as in Root nodes, as basic units of description makes it possible to characterize a variety of language patterns in which gestural organization varies. Such patterns range from the misorderings of disordered speech through phonological rules involving gestural overlap and deletion to historical changes in which the overlap of gestures provides a crucial explanatory element.

Gestures can participate in language patterns involving overlap because they are spatiotemporal in nature and therefore have internal duration. In addition, gestures differ from current theories of feature geometry by including the constriction degree as an inherent part of the gesture. Since the gestural constrictions occur in the vocal tract, which can be characterized in terms of tube geometry, all the levels of the vocal tract will be constricted, leading to a constriction degree hierarchy. The values of the constriction degree at each higher level node in the hierarchy can be predicted on the basis of the percolation principles and tube geometry. In this way, the use of gestures as atoms can be reconciled with the use of constriction degree at various levels in the vocal tract (or feature geometry) hierarchy.

The phonological notation developed for the gestural approach might usefully be incorporated, in whole or in part, into other phonologies. Five components of the notation were discussed, all derived from the basic premise that gestures are the primitive phonological unit, organized into gestural scores. These components include (1) constriction degree as a subordinate of the articulator node and (2) stiffness (duration) as a subordinate of the articulator node. That is, both CD and duration are inherent to the gesture. The gestures are arranged in gestural scores using (3) articulatory tiers, with (4) the relevant geometry (articulatory, tube or feature) indicated to the left of the score and (5) structural information above the score, if desired. Association lines can also be used to indicate how the gestures are combined into phonological units. Thus, gestures can serve both as characterizations of articulatory movement data and as the atoms of phonological representation.

REFERENCES

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Abry, C., & Bõe, L.-J. (1986). 'Laws' for lips. *Speech Communication, 5*, 97-104.

Anderson, J., & Ewen, C. J. (1987). *Principles of dependency phonology*. Cambridge: Cambridge University Press.

Anderson, S. R. (1974). *The organization of phonology*. New York: Academic Press.

Barry, M. C. (1985). A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics, 4*, 1-16.

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Best, C. T., & D. Wilkenfeld (1988). Phonologically motivated substitutions in a 20-22 month old's imitations of intervocalic alveolar stops. *Journal of the Acoustical Society of America, 84,* (S1), S114. (Abstract)

Boucher, V. (1988). For a phonology without 'rules': Peripheral timing-structures and allophonic patterns in French and English. In S. Embleton (Ed.), *The Fourteenth LACUS Forum 1987* (pp. 123-140). Lake Bluff IL: Linguistic Association of Canada and the U.S.

Boysson-Bardies, B. de, Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 11,* 1-15.

Boysson-Bardies, B. de, Sagart, L., Halle, P., & Durand, C. (1986). Acoustic investigations of cross-linguistic variability in babbling. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 113-126). Basingstoke, Hampshire: MacMillan.

Browman, C. P., & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In V. Fromkin (Ed.), *Phonetic linguistics* (pp. 35-53). New York: Academic.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3,* 219-252.

Browman, C. P., & Goldstein, L. (1987). Tiers in Articulatory Phonology, with some implications for casual speech. *Haskins Laboratories Status Report on Speech Research, SR-92* (pp. 1-30). To appear in J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech.* Cambridge: Cambridge University Press.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica, 45,* 140-155.

Browman, C. P., & Goldstein, L. (1989). 'Targetless' schwa: An articulatory analysis. Paper presented at the Second Conference on Laboratory Phonology. Edinburgh, June 29-July 2, 1989.

Browman, C. P., & Goldstein, L. (in press). Gestural structures and phonological patterns. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Browman, C. P., Goldstein, L., Kelso, J. A. S., Rubin, P., & Saltzman, E. (1984). Articulatory synthesis from underlying dynamics. *Journal of the Acoustical Society of America, 75,* S22-S23. (Abstract)

Browman, C. P., Goldstein, L., Saltzman, E., & Smith, C. (1986). GEST: A computational model for speech production using dynamically defined articulatory gestures. *Journal of the Acoustical Society of America, 80,* S97. (Abstract)

Brown, G. (1977). *Listening to spoken English.* London: Longman.

Campbell, L. (1974). Phonological features: Problems and proposals. *Language, 50,* 52-65.

Catford, J. C. (1977). *Fundamental problems in phonetics.* Bloomington: Indiana University Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper and Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook, 2,* 225-252.

Clements, G. N., & Keyser, S. J. (1983). *CV phonology: A generative theory of the syllable.* Cambridge, MA: MIT Press.

Ewen, C. J. (1982). The internal structure of complex segments. In H. van der Hulst & N. Smith (Eds.), *The structure of phonological representations* (pp. 27-67). Cinnaminson, NJ: Foris Publications U.S.A.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51,* 419-439.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8,* 113-133.

Fowler, C. A. (1981). A relationship between coarticulation and and compensatory shortening. *Phonetica, 38,* 35-50.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet. *Journal of Experimental Psychology: General, 112,* 386-412.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production* (pp. 373-420). New York: Academic Press.

Fry D. B. (1966). The development of the phonological system in the normal and the deaf child. In F. Smith & G. Miller (Eds.), *The genesis of language: A psycholinguistic approach* (pp. 187-206). Cambridge, MA: MIT Press.

Fujimura, O. (1981a). Temporal organization of articulatory movements as a multidimensional phrasal structure. *Phonetica, 38,* 66-83.

Fujimura, O. (1981b). Elementary gestures and temporal organization—What does an articulatory constraint mean? In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 101-110). Amsterdam: North Holland.

Gimson, A. C. (1962). *An introduction to the pronunciation of English.* London: Edward Arnold Publishers, Ltd.

Glass, L., & Mackey, M. C. (1988). *From clocks to chaos.* Princeton: Princeton University Press.

Goldstein, L., & Browman, C. P. (1986). Representation of voicing contrasts using articulatory gestures. *Journal of Phonetics, 14,* 339-342.

Gregores, E., & Suárez, J. A. (1967). *A description of colloquial Guaraní.* The Hague: Mouton.

Guy, G. R. (1980). Variation in the group and the individual: The case of final stop deletion. In W. Labov (Ed.), *Locating language in time and space* (pp. 1-36). New York: Academic Press.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51,* 347-356.

Halle, M. (1982). On distinctive features and their articulatory implementation. *Natural Language and Linguistic Theory, 1,* 91-105.

Halle, M., & Stevens, K. (1971). A note on laryngeal features. *MIT-RLE Quarterly Progress Report, 11,* 198-213.

Hammond, M. (1988). On deriving the well-formedness condition. *Linguistic Inquiry, 19,* 319-325.

Hardcastle, W. J., & Roach, P. J. (1979). An instrumental investigation of coarticulation in stop consonant sequences. In P. Hollien & H. Hollien (Eds.), *Current issues in the phonetic sciences (pp. 531-540).* Amsterdam: John Benjamins AG.

Hayes, B. (1986). Assimilation as spreading in Toba Batak. *Linguistic Inquiry, 17,* 467-499.

Jakobson, R., Fant, C. G. M., & Halle, M. (1969). *Preliminaries to speech analysis: The distinctive features and their correlates.* Cambridge, MA: The MIT Press.

Jespersen, O. (1914). *Lehrbuch der phonetik.* Leipzig: B. G. Teubner.

Keating, P. A. (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA Working Papers in Phonetics, 62,* 1-13.

Keating, P. A. (1988). Palatals as complex segments: X-ray evidence. *UCLA Working Papers in Phonetics, 69,* 77-91.

Kelso, J. A. S., & Tuller, B. (1987). Intrinsic time in speech production: Theory, methodology, and preliminary observations. In E. Keller & M. Gopnik (Eds.), *Motor and sensory processes of language* (pp. 203-222). Hillsdale, NJ: Erlbaum.

Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech

production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77*, 266-280.

Kent, R. D. (1983). The segmental organization of speech. In P.F. MacNeilage (Ed.), *The production of speech* (pp. 57-89). New York Springer-Verlag.

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America, 72*, 353-365.

Kohler, K. (1976). Die Instabilität wortfinaler Alveolarplosive im Deutschen: Eine elektropalatographische Untersuchung [The instability of word-final alveolar plosives in German: An electropalatographic investigation.] *Phonetica, 33*, 1-30.

Koopmans-van Beinum, F. J., & Van der Stelt, J. M. (1986). Early stages in the development of speech movements. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 37-50). Basingstoke, Hampshire: MacMillan.

Krakow, R. A. (1989). *The articulatory organization of syllables: A kinematic analysis of labial and velar gestures*. Doctoral dissertation, Yale University.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56*, 485-502.

Ladefoged, P. (1982). *A course in phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

Ladefoged, P. (1988a). Hierarchical features of the International Phonetic Alphabet. *UCLA Working Papers in Phonetics, 70*, 1-12.

Ladefoged, P. (1988b). The many interfaces between phonetics and phonology. *UCLA Working Papers in Phonetics, 70*, 13-23.

Ladefoged, P., & Halle, M. (1988). Some major features of the International Phonetic Alphabet. *Language, 64*, 577-582.

Ladefoged, P., & Maddieson, I. (1986). Some of the sounds of the world's languages. *UCLA Working Papers in Phonetics, 64*, 1-137.

Lass, R. (1984). *Phonology: An introduction to basic concepts*. Cambridge: Cambridge University Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Lieberman, P. (1984). *The biology and evolution of language*. Cambridge: Harvard University Press.

Lindau, M. (1978). Vowel features. *Language, 54*, 541-563.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. Jaeger (Eds.), *Experimental phonology* (pp. 13-44). Orlando: Academic Press.

Lindblom, B., MacNeilage, P. F., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie & Ö. Dahl (Eds.), *Explanations of linguistic universals* (pp. 182-203). The Hague: Mouton.

Linker, W. (1982). Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study. *UCLA Working Papers in Phonetics, 56*, 1-134.

Locke, J. L. (1983). *Phonological acquisition and change*. New York: Academic Press.

Locke, J. L. (1986). The linguistic significance of babbling. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 143-160). Basingstoke, Hampshire: MacMillan.

Lombardi, L. (1987). On the representation of the affricate. Ms, University of Massachusetts, Amherst.

Lynch, J. (1983). On the Kuman liquids. *Language and Linguistics in Melanesia, 14*, 98-112.

Maddieson, I. (1987). Revision of the IPA: Linguo-labials as a test case. *Journal of the International Phonetic Association, 17*, 26-30.

Maddieson, I., & Ladefoged, P. (1989). Multiply articulated segments and the feature hierarchy. *UCLA Working Papers in Phonetics, 72*, 116-138.

Marshall, R. C., Gandour, J., & Windsor, J. (1988). Selective impairment of phonation: A case study. *Brain and Language, 35*, 313-339.

McCarthy, J. J. (1988). Feature geometry and dependency. *Phonetica, 45*, 84-108

McCarthy, J. J. (1989). Linear ordering in phonological representation. *Linguistic Inquiry, 20*, 71-99.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research, 32*, 120-132.

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America, 84*, 1653-1661.

Ohala, J. J. (1975). Phonetic explanations for nasal sound patterns. In C. A. Ferguson, L. M. Hyman, & J. J. Ohala (Eds.), *Nasálfest* (pp. 289-316). Stanford: Stanford University.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178-203). Chicago: Chicago Linguistic Society.

Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 189-216). New York: Springer-Verlag.

Ohala, J. J. (1985). Around 'flat.' In V. A. Fromkin (Ed.), *Phonetic Linguistics: Essays in honor of Peter Ladefoged* (pp. 223-241). New York: Academic Press.

Ohala, J. J., & Lorentz, J. (1977). The story of [w]: An exercise in the phonetic explanation for sound patterns. *Proceedings, Annual Meeting of the Berkeley Linguistic Society, 3*, 577-599.

Oller, D. K. (1986). Metaphonology and infant vocalizations. In B. Lindblom & R. Zetterström (Eds.), *Precursors of early speech* (pp. 21-36). Basingstoke, Hampshire: MacMillan.

Oller, D. K., & Eilers, R. E. (1982). Similarity of babbling in Spanish- and English-learning babies. *Journal of Child Language, 9*, 565-577.

Oller, D. K. & Eilers, R. E. (1988). The role of audition in infant babbling. *Child Development, 59*, 441-449.

Ostry, D. J. & Munhall, K. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*, 640-648.

Pagliuca, W. (1982). *Prolegomena to a theory of articulatory evolution*. Doctoral dissertation, State University of New York at Buffalo. Ann Arbor, MI: University Microfilms International.

Pike, K. L. (1943). *Phonetics*. Ann Arbor: University of Michigan Press.

Plotkin, V. Y. (1976). Systems of ultimate units. *Phonetica, 33*, 81-92.

Principles of the International Phonetic Association (1949, reprinted 1978). London University College.

Recasens, D. (in preparation). The articulatory characteristics of alveolopalatal and palatal consonants. Haskins Laboratories. New Haven, CT.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70*, 321-328.

Sagey, E. C. (1986). *The representation of features and relations in non-linear phonology*. Unpublished doctoral dissertation, MIT.

Sagey, E. C. (1988). On the ill-formedness of crossing association lines. *Linguistic Inquiry, 19*, 109-118.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators A preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (pp. 129-144). *(Experimental Brain Research Series 15)*. New York Springer-Verlag.

Saltzman, E., Goldstein, L., Browman, C. P., & Rubin, P. E. (1988). Dynamics of gestural blending during speech production.

Presented at 1st Annual International Neural Network Society (INNS) Boston, September 6-10, 1988.

Saltzman, E., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review, 94*, 84-106.

Saltzman, E., & K. Munhall (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*.

Saltzman, E., Rubin, P. E., Goldstein, L., & Browman, C. P. (1987). Task-dynamic modeling of interarticulator coordination. *Journal of the Acoustical Society of America, 82*, S15. (Abstract)

Schadeburg, T. C. (1982). Nasalization in UMbundu. *Journal of African Languages and Linguistics, 4*, 109-132.

Schmidt, R. C., Carello, C., & Turvey, M. T. (1987). Visual coupling of biological oscillators. *PAW Review, 2(1)*, 22-24.

Selkirk, E. (1988). A two-root theory of length. Presented at the 19th Meeting of the New England Linguistic Society (NELS), November 1988.

Shockey, L. (1974). Phonetic and phonological properties of connected speech. *Ohio State Working Papers in Linguistics, 17*, iv-143.

Steriade, D. (in press). Gestures and autosegments: Comments on Browman and Goldstein's 'Gestures in Articulatory Phonology.' In J. Kingston & M. E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51-66). New York McGraw Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 113-45.

Stevens, K. N., Keyser, S. J., & Kawasaki, H. (1986). Toward a phonetic and phonological theory of redundant features. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 426-449). Hillsdale, NJ: Lawrence Erlbaum Associates.

Straight, H. S. (1976). *The acquisition of Maya phonology: Variation in Yucatec child language*. New York Garland.

Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz & E. Scheerer (Eds.), *Language perception and production* (pp. 67-84). London Academic Press.

Thelen, E. (1981). Rhythmical behavior in infancy: An ethological perspective. *Developmental Psychology, 17*, 237-257.

Thompson, J. M. T., & Stewart, H. B. (1976). *Nonlinear dynamics and chaos*. New York: John Wiley & Sons.

Thráinsson, H. (1978). On the phonology of Icelandic preaspiration. *Nordic Journal of Linguistics, 1(1)*, 3-54.

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics*. Bloomington: Indiana Linguistics Club.

Vennemann, T., & Ladefoged, P. (1973). Phonetic features and phonological features. *Lingua, 32*, 61-74.

Vihman, M. M. (in press). Ontogeny of phonetic gestures: Speech production. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., & Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. *Language, 61*, 397-445.

Wood, S. (1982). X-ray and model studies of vowel articulation. *Working Papers in Linguistics, 23*, Lund University.

FOOTNOTES

* *Phonology, 6*, 201-151 (1989).

† Also, Department of Linguistics, Yale University.

[1] A simple dynamical system consists of a mass attached to the end of a spring—a damped mass-spring model. If the mass is pulled, stretching the spring beyond its rest length (equilibrium position), and then released, the system will begin to oscillate. The resultant movement patterns of the mass will be a damped sinusoid described by the solution to the equation below. When such an equation is used to model the movements of coordinated sets of articulators, the 'object'—motion variable—in the equation is considered to be the tract variable, for example, lip aperture (LA). Thus, the sinusoids trajectory would describe how lip aperture changes over time:

$$m\ddot{x} + b\dot{x} + k(x - x0) = 0$$

where m = mass of the object, b = damping of the system, k = stiffness of the spring, $x0$ = rest length of the spring (equilibrium position), x = instantaneous displacement of the object, $\dot{x}$ = instantaneous velocity of the object, $\ddot{x}$ = instantaneous acceleration of the object.

[2] For ease of reference to the gesture, it is possible to use either a bundle of gestural descriptors (or a selected subset) or gestural symbols. We have been trying different approaches to the question of what gestural symbols should be; our present best estimate is that gestures should be treated like archiphonemes. Thus, our current proposal is to use the capitalized form of the voiced IPA symbol for oral gestures, capitalized and diacritized {H} for glottal gestures, and {±N} for [velic clo] and [velic open] gestures, respectively. In order to clearly distinguish gestural symbols from other phonetic symbols, we enclose them in curly brackets: { }. This approach should permit gestural descriptions to draw upon the full symbol resources of IPA, rather than attempting to develop an additional set of symbols. However, we welcome comments from others on this decision, particularly on the proposal to capitalize gestural symbols, a choice that minimizes confusion with phonemic transcriptions—but that leads to a conflict with uvular symbols in the current IPA system.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 102-117

The Perception of Phonetic Gestures*

Carol A. Fowler† and Lawrence D. Rosenblum††

We have titled our presentation "The perception of phonetic gestures" as if phonetic gestures *are* perceived. By phonetic gestures we refer to organized movements of one or more vocal-tract structures that realize phonetic dimensions of an utterance (cf. Browman & Goldstein, 1986; in press a). An example of a gesture is bilabial closure for a stop, which includes contributions by the jaw and the upper and lower lips. Gestures are organized into larger segmental and suprasegmental groupings, and we do not intend to imply that these larger organizations are not perceived as well. We focus on gestures to emphasize a claim that, in speech, perceptual objects are fundamentally articulatory as well as linguistic.

That is, in speech perception, articulatory events have a status quite different from that of their acoustic products. The former are perceived, whereas the latter are the means (or one of the means) by which they are perceived.

A claim that phonetic gestures are perceived is not uncontroversial, of course, and there are other points of view (e.g., Massaro, 1987; Stevens & Blumstein, 1981). We do not intend to consider these other views here, however, but instead to focus on agreements and disagreements between two theoretical perspectives from which the claim is made. Accordingly, we begin by summarizing some of the evidence that, in our view, justifies it.

Phonetic gestures are perceived: Three sources of evidence

1. Correspondence failures between acoustic signal and percept: Correspondences between gestures and percept

Perhaps the most compelling evidence that gestures, and not their acoustic products, are perceptual objects is the failure of dimensions of speech percepts to correspond to obvious dimensions of the acoustic signal and their correspondence, instead, to phonetically-organized articulatory behaviors that produce the signal. We offer three examples, all of them implicating articulatory gestures as perceptual objects and the third showing most clearly that the perceived gestures are not surface articulatory movements, but rather, linguistically-organized gestures.

a. Synthetic /di/ and /du/

One example from the early work at Haskins Laboratories (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) is of synthetic /di/ and /du/. Monosyllables, such as those in Figure 1, can be synthesized that consist only of two formants. The information specifying /d/ (rather than /b/ or /g/) in both syllables is the second formant transition. These transitions are very different in the two syllables, and, extracted from their syllables, they sound very different too. Each sounds more-or-less like the frequency glide it resembles in the visible display. Neither sounds like /d/. In the context of their respective syllables, however, they sound alike and they sound like /d/.

The consonantal segments in /di/ and /du/ are produced alike too, by a constriction and release gesture of the tongue tip against the alveolar ridge of the palate. When listeners perceive the

Preparation of this manuscript was supported by a fellowship from the John Simon Guggenheim Memorial Foundation to the first author and by grants NIH-NICHD HD-01994 and NINCDS NS-13617 to Haskins Laboratories.

synthetic /di/ and /du/ syllables of Figure 1, their percepts correspond to the implied constriction and release gestures, not, it seems, to the context-sensitive acoustic signal.

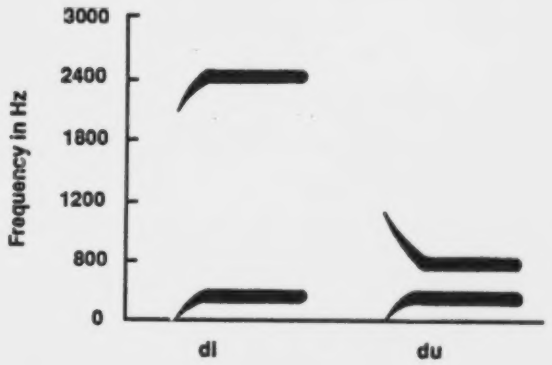


Figure 1. Synthetic syllables /di/ and /du/. The second formant transitions identify the initial consonant as /d/ rather than as /b/ or /g/.

b. Functional equivalence of acoustic "cues"

We expect listeners to be very good at distinguishing an interval of silence from nonsilence—from a set of frequency glides, for example. Too, we expect them to distinguish acoustic signals that differ in two ways more readily than signals that differ in just one of the two ways. Both of these expectations are violated—another example of noncorrespondence—if the silence and glides are joint acoustic products of a common constriction and release gesture for a stop consonant.

Fitch, Halwes, Erickson and Liberman (1980) created synthetic syllables identified as "slit" or "split" by varying the duration of a silent interval following the fricative and manipulating the presence or absence of transitions for a bilabial stop following the silent interval. A relatively long silent interval and the presence of transitions both signal a bilabial stop, the silent interval cuing the closure and the transitions the release. Fitch et al. found that pairs of syllables differing on both cue dimensions, duration of silence and presence/absence of transitions, were either more discriminable than pairs differing in one of these ways or less discriminable depending on how the cues were combined. A syllable with a long silent interval and transitions was highly discriminable from a syllable with a shorter silent interval and no transitions; the one was identified as "split" and the other as "slit." A syllable with a short silent interval and transitions was nearly indiscriminable from one with a longer interval and no transitions; both were identified as "split." Syllables differing in two ways are indiscriminable just when the acoustic cues that distinguish them are "functionally equivalent"—that is, they cue the same articulatory gesture. A long silent interval does not normally sound like a set of frequency glides, but it does in a context in which each specifies a consonantal constriction.

c. Perception of intonation

The findings just summarized, among others, reveal that listeners perceive gestures. Apparently, listeners do not perceive the acoustic signal *per se*.

Nor, however, do they perceive "raw" articulatory motions as such. Rather, they perceive linguistically-organized (phonetic) gestures. Research on the various ways in which fundamental frequency (henceforth, f_0) is perceived shows this most clearly.

Perceived intonational peak height will not, in general, correspond to the absolute rate at which the vocal folds open and close during production of the peak. Instead, perception of the peak corresponds to just those influences on the rate of opening and closing that are caused by gestures intended by the talker to affect intonational peak height. (Largely, intonational melody is implemented by contraction and relaxation of muscles of the larynx that tense the vocal folds; see, e.g., Ohala, 1978.) There are other influences on the rate of vocal fold opening and closing that may either decrease or increase f_0. Some of these influences, due to lung deflation during an expiration ("declination," Gelfer, Harris, & Baer, 1987; Gelfer, Harris, Collier, & Baer, 1985) or to segmental perturbations reflecting vowel height (e.g., Lehiste & Peterson, 1961) and obstruent voicing (e.g., Ohde, 1984), are largely or entirely automatic consequences of other things that talkers are doing (producing an utterance on an expiratory airflow, producing a close or open vowel [Honda, 1981], producing a voiced or voiceless obstruent [Ohde, 1984]; Löfqvist, Baer, McGarr, & Seider Story, in press). They do not sound like changes in pitch; rather, they sound like what they are: information for early-to-late serial position in an utterance in the case of declination (Pierrehumbert, 1979; see also Lehiste, 1982), and information for vowel height (Reinholt Peterson, 1986; Silverman, 1987) or consonant voicing (e.g., Silverman, 1986) in the case of segmental perturbations.

As we will suggest below (under "How acoustic structure may serve as information for speech perceivers"), listeners apparently use configurations of changes in different acoustic variables to recover the distinct, organized articulatory systems that implement the various linguistic dimensions of talkers' utterances. By using acoustic information in this way, listeners can recover what Liberman (1982) has called the talker's "phonetic intents."

2. Audio-visual integration of gestural information

A video display of a face mouthing /ga/ synchronized with an acoustic signal of the speaker saying /ba/ is heard most typically as "da" (MacDonald & McGurk, 1978). Subjects' identifications of syllables presented in this type of experiment reflect an integration of information from the optical and acoustic sources. Too, as Liberman (1982) points out, the integration affects what listeners experience *hearing* to an extent that they cannot tell what contribution to their perceptual experience is made by the acoustic signal and what by the video display.[1]

Why does integration occur? One answer is that both sources of information, the optical and the acoustic, provide information apparently about the same event of talking, and they do so by providing information about the talkers' phonetic gestures.

3. Shadowing

Listeners' latency to repeat a syllable they hear is very short—in Porter's research (Porter, 1978; Porter & Lubker, 1980), around 180 ms on average. Even though these latencies are obtained in a choice reaction time procedure (in which the vocal response required is different for different stimuli to respond), latencies approach simple reaction times (in which the same response occurs to any stimulus to respond), and they are much shorter than choice reaction times using a button press.

Why should these particular choice reaction times be so fast? Presumably, the compatibility between stimulus and response explains the fast response times. Indeed, it effectively eliminates the element of choice. If listeners perceive the talker's phonetic gestures, then the only response requiring essentially no choice at all is one that reproduces those gestures.

The motor theory

Throughout most of its history, the motor theory (e.g., Liberman, Cooper, Harris, & MacNeilage, 1963; Liberman et al., 1967; Liberman & Mattingly, 1985; see also, Cooper, Delattre, Liberman, Borst, & Gerstman, 1952) has been the only theory of speech perception to identify the phonetic gesture as an object of perception. Here we describe the motor theory by discussing what, more precisely, the motor theorists have considered to be the object of perception, how they characterize the process of speech perception and why, recently, they have introduced the idea that speech perception is accomplished by a specialized module.

What is perceived for the motor theorist?

Coarticulation is the reason why the acoustic signal appears to correspond so badly to the sequences of phonemes that talkers intend to produce. Due to coarticulation, phonemes are produced in overlapping time frames so that the acoustic signal is everywhere (or nearly everywhere; see, e.g., Stevens & Blumstein, 1981), context-sensitive. This makes the signal a complex "code" on the phonemes of the language, not a cipher, like an alphabet.[2] In "Perception of the speech code" (1967), Liberman and his colleagues speculated that coarticulatory "encoding" is, in part, a necessary consequence of properties of the speech articulators (their sluggishness, for example). However, in their view, coarticulation is also promoted both by the nature of phonemes themselves—that they are realized by sets of subphonemic features[3]— and by the listener's short-term memory, which would be overtaxed by the slow transmission rate of an acoustic cipher.

In producing speech, talkers exploit the fact that the different articulators—the lips, velum, jaw, etc.—can be independently controlled. Subphonemic features, such as lip rounding, velum lowering and alveolar closure each use subsets of the articulators, often just one; therefore, more than one feature can be produced at a time. Speech can be produced at rapid rates by allowing "parallel transmission" of the subphonemic features of different phonemes. This increases the transmission rates for listeners, but it also creates much of the encoding that is considered responsible for the apparent lack of invariance between acoustic and phonetic segments.

The listener's percept corresponds, it seems, neither to the encoded cues in the acoustic signal nor even to the also-encoded succession of vocal tract shapes during speech production, but instead to a sequence of discrete, unencoded phonemes, each composed of its own component subphonemic

features. To explain why "perception mirrors articulation more closely than sound" (p. 453) and (yet) achieves recovery of discrete unencoded phonemes, the motor theorists proposed as a first hypothesis that perceivers somehow access their speech-motor systems in perception and that the percept they achieve corresponds to a stage in production before encoding of the speech segments takes place. In "Perception of the speech code," the stage was one in which "motor commands" to the muscles were selected to implement subphonemic features. In "The motor theory revised," (Liberman & Mattingly, 1985), a revision to the theory reflects developments in our understanding of motor control. Evidence suggests that activities of the vocal tract are products of functional couplings among articulators (e.g., Folkins & Abbs, 1975, 1976; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984), which produce gestures as defined earlier, not independent movements of the articulators identified with subphonemic features in "Perception of the speech code." In "The motor theory revised," control structures for gestures have replaced motor commands for subphonemic features as invariants of production and as objects of perception for listeners.[4] Like subphonemic features, control structures are abstract, prevented by coarticulation from making public appearances in the vocal tract. Liberman and Mattingly write of the perceptual objects of the revised theory:

> We would argue, then, that the gestures do have characteristic invariant properties, as the motor theory requires, though these must be seen, not as peripheral movements, but as the more remote structures that control the movements. These structures correspond to the speaker's intentions. (p. 23)

In recovering abstract gestures, processes of speech perception yield quite different kinds of perceptual objects than general auditory perception. In auditory perception, more generally, according to Liberman and Mattingly, listeners hear the signal as "ordinary sound" (p. 6); that is, they hear the acoustic signal as such. In other publications, Mattingly and Liberman (1988) refer to this apparently more straightforward perceptual object as "homomorphic" in contrast to objects of speech perception which are "heteromorphic." An example they offer of homomorphic auditory perception is perception of isolated formant transitions which sound like the frequency glides they resemble in a spectrographic display.

How perception takes place in the motor theory

In the motor theory, listeners use "analysis by synthesis" to recover phonetic gestures from the encoded, informationally-impoverished acoustic signal. This aspect of the theory has never been worked out in detail. However, in general, analysis by synthesis consists in analyzing a signal by guessing how the signal might have been produced (e.g., Stevens, 1960; Stevens & Halle, 1964). Liberman and Mattingly refer to an "internal, innately specified vocal-tract synthesizer that incorporates complete information about the anatomical and physiological characteristics of the vocal tract and also about the articulatory and acoustic consequences of linguistically significant gestures" (p. 26). The synthesizer computes candidate gestures and then determines which of those gestures, in combination with others identified as ongoing in the vocal-tract could account for the acoustic signal.

Speech perception as modular

If speech perception does involve accessing the speech-motor system, then it must indeed be special and quite distinct from general auditory perception. It is special in its objects of perception, in the kinds of processes applied to the acoustic signal, and presumably in the neural systems dedicated to those processes as well. Liberman and Mattingly propose that speech perception is achieved by a specialized module.

A module (Fodor, 1983) is a cognitive system that tends to be narrowly specialized ("domain specific"), using computations that are special ("eccentric") to its domain; it is computationally autonomous (so that different systems do not compete for resources) and prototypically is associated with a distinct neural substrate. In addition, modules tend to be "informationally encapsulated," bringing to bear on the processing they do only some of the relevant information the perceiver may have; in particular, processing of "input" (perceptual) systems—prime examples of modules—is protected early on from bias by "top-down" information.

The speech perceptual system of the motor theory has all of these characteristics. It is narrowly specialized and its perception-production link is eccentric; moreover, it is associated with a specialized neural substrate (e.g., Kimura, 1961). In addition, as the remarkable phenomenon of duplex perception (e.g., Liberman, Isenberg, & Rakerd, 1981; Rand, 1974) suggests, the speech

perceiving system is autonomous and informationally encapsulated.

In duplex perception as it is typically investigated (e.g., Liberman et al., 1981; Mann & Liberman, 1983; Repp, Milburn, & Ashkenas, 1983), most of an acoustic CV syllable (the "base" at the left of Figure 2) is presented to one ear while the remainder, generally a formant transition (either of the "chirps" on the right side of Figure 2) is presented to the other ear. Heard in isolation, the base is ambiguous between "da" and "ga," but listeners generally report hearing "da." (It was identified as "da" 87% of the time in the study by Repp et al., 1983.) In isolation, the chirps sound like the frequency glides they resemble; they do not sound speech-like. Presented dichotically, listeners integrate the chirp and the base, hearing the integrated "da" or "ga" in the ear receiving the base. Remarkably, in addition, they hear the chirp in the other ear. Researchers who have investigated duplex perception describe it as perception of the same part of an acoustic signal in two ways simultaneously. If that characterization is correct, it implies strongly that the percepts are outputs of two distinct and autonomous perceptual systems, one specialized for speech and the other perhaps general to other acoustic signals.

A striking characteristic of speech perceptual systems that integrate syllable fragments presented to different ears is their imperviousness to information in the spatial separation of the fragments that they cannot possibly be part of the same spoken syllable—an instance, perhaps, of information encapsulation.

In recent work, Mattingly and Liberman (in press) have revised, or expanded on, Fodor's view of modules by proposing a distinction between "closed" and "open" modules. Closed modules, including the speech module and a sound-localization module, for example, are narrowly specialized as Fodor has characterized modules more generally. In addition (among other special properties), they yield heteromorphic percepts—that is, percepts whose dimensions are not those of the proximal stimulation. Although Mattingly and Liberman characterize the heteromorphic percept in this way—in terms of what it does *not* conform to, it appears that the heteromorphic percept can be characterized in a more positive way as well. The dimensions of heteromorphic percepts are those of distal events, not of proximal stimulation. The speech module renders phonetic gestures; the sound-localization module renders location in space. By contrast, open modules are sort of "everything-else" perceptual systems.

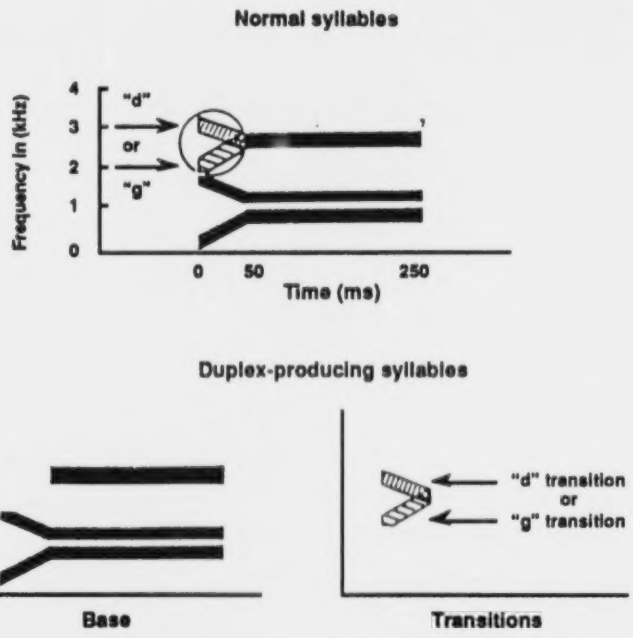


Figure 2. Stimuli that yield duplex perception. The base is presented to one ear and the third formants to another. In the ear to which the base is presented, listeners hear the syllable specified jointly by the base and the transitions; in the other ear, they hear the transitions as frequency glides. (Figure adapted from Whalen & Liberman, 1987).

An open auditory-perceptual module is responsible for perception of most sounds in the environment. According to the theory, outputs of open modules are homomorphic.

In the context of this account of auditory perception, the conditions under which duplex perception is studied are seen as somehow tricking the open module into providing a percept of the isolated formant transition even though the transition is also being perceived by the speech module. Accordingly, two percepts are provided for one acoustic fragment; one percept is homomorphic and the other is heteromorphic.

Prospectus

Our brief overview of the motor theory obviously cannot do justice to it. In our view, it is, to date, superior to other theories of speech perception in at least two major respects: in its ability to handle the full range of behavioral findings on speech perception—in particular, of course, the evidence that listeners recover phonetic gestures—and in having developed its account of speech in the context of a more general theory of biological specializations for perception.

Our purpose here, however, is not just to praise the theory, but to challenge it as well, with the further aim of provoking the motor theorists either to buttress their theory where it appears to us vulnerable, or else to revise it further.

We will raise three general questions from the perspective of our own, direct-realist theory (Fowler, 1986a,b; Rosenblum, 1987). First we question the inference from evidence that listeners recover phonetic gestures that the listener's own speech-motor system plays a role in perception. The nature of the challenge we mount to this inference leads to a second one. We question the idea that, whereas a specialized speech module—and other closed modules—render heteromorphic percepts, other percepts are homomorphic. Finally, we challenge the idea in any case that duplex perception reveals that speech perception is achieved by a closed module.

Standing behind all of these specific questions we raise about claims of the motor theory is a general issue that needs to be confronted by all of us who study speech perception and, for that matter, perception more generally. The issue is one of determining when behavioral data warrant inferences being drawn about perceptual processes taking place inside perceivers and when the data deserve accounting instead in terms of the nature of events taking place publicly when something is perceived.

Does perceptual recovery of phonetic gestures implicate the listener's speech motor system?

In our view, the evidence that perceivers recover phonetic gestures in speech perception is incontrovertible[5] and any theory of speech perception is inadequate unless it can provide a unified account of those findings. However, the motor theorists have drawn an inference from these findings that, we argue, is not warranted by the general observation that listeners recover gestures. The inference is that recovery of gestures implies access by the perceiver to his own speech-motor system. It is notable, perhaps, that, in neither "Perception of the speech code" nor "The motor theory revised," do Liberman and his colleagues offer any evidence in support of this claim except evidence that listeners recover gestures (and that human left-cerebral hemispheres are specialized for speech and especially for phonetic perception [e.g., Kimura, 1961; Liberman, 1974; Studdert-Kennedy & Shankweiler, 1970]).

There is another way to explain why listeners recover phonetic gestures. It is that phonetic gestures are among the "distal events" that occur when speech is perceived and that perception universally involves recovery of distal events from information in proximal stimulation.

Distal events universally are perceptual objects: Proximal stimuli universally are not.

Consider first visual perception observed from outside the perceiver.[6] Visual perceivers recover properties of objects and events in their environment ("distal events"). They can do so, in part, because the environment supplies information about the objects and events in a form that their perceptual systems can use. Light reflects from objects and events, which structure it lawfully; given a distal event and light from some source, the reflected light must have the structure that it has. To the extent that the structure in the light is also specific to the properties of a distal event that caused it, it can serve as information to a perceiver about its distal source. The reflected light ("proximal stimulation") has another property that permits it its central role in perception. It can stimulate the visual system of a perceiver and thereby impart its structure to it. From there, the perceiver can use the structure as information for distal-event perception.

The reflected light does not provide information to the visual system by picturing the world. Information in reflected light for "looming" (that is, for an object on a collision course with the perceiver's head), for example, is a certain manner of expansion of the contours of the object's reflection in the light, that progressively covers the contours of optical reflections of immobile parts of the perceiver's environment. When an object looms, it does not grow; it approaches. However, its optical reflection grows, and, confronted with such an optic array, perceivers (from fiddler crabs to kittens to rhesus monkeys to humans [Schiff, 1965; Schiff, Caviness, & Gibson, 1962]) behave as if they perceive an object on a collision course; that is, they try to avoid it.

Two related conclusions from this characterization of visual perception are first that observers see distal events based on information about them in proximal stimulation and second that, in Mattingly and Liberman's terms, visual perception therefore is quite generally heteromorphic. It is not merely heteromorphic in respect to those aspects of stimulation handled by closed modules (for example, one that recovers depth information from binocular disparity); it is *generally* the case that the dimensions of the percept correspond with dimensions of distal objects and events and not necessarily with those of a distal-event-free description of the proximal stimulation.[7]

Auditory perception is analogous to visual perception in its general character, viewed, once again from outside the perceiver. Consider any sounding object, a ringing bell, for example. The ringing bell is a "distal event" that structures an acoustic signal. The structuring of the air by the bell is lawful and, to the extent that it also tends to be specific to its distal source, the structure can provide information about the source to a sensitive perceiver. Like reflected light, the acoustic signal (the proximal stimulation) in fact has two critical properties that allow it to play a central role in perception. It is lawfully structured by some distal event and it can stimulate the auditory system of perceivers, thereby imparting its structure to it. The perceiver then can use the structure as information for its source.

As for structure in reflected light, structure in an acoustic signal does not resemble the sound-producing source in any way. Accordingly, if auditory perception works similarly to visual perception—that is, if perceivers use structure in acoustic signals to recover their distal sources, then auditory percepts, like visual percepts will be heteromorphic.

Liberman and Mattingly (1985; Mattingly & Liberman, 1988) suggest, however, that in general, auditory perceptions are homomorphic. We agree that our intuitions are less clear here than they are in the case of visual perception. However, it is an empirical question whether dimensions of listeners' percepts are better explained in terms of dimensions of distal events or of a distal-event free description of proximal stimulation. To date the question is untested, however; for whatever reason, researchers who study auditory perception rarely study perception of natural sound-producing events (see, however, Repp, 1987; VanDerVeer, 1979; Warren & Verbrugge, 1984).

Now consider speech perception. In speech, the distal event—at least the event in the environment that structures the acoustic speech signal—is the moving vocal tract. If, as we propose, the vocal tract produces phonetic gestures, then the distal event is, at the same time, the set of phonetic gestures that compose the talker's spoken message. The proximal stimulus is the acoustic signal, lawfully structured by movement in the vocal tract. To the extent that the structure in the signal also tends to be specific to the events that caused it, it can serve as information about those events to sensitive perceivers. The information that proximal stimulation provides will be about the phonetic gestures of the vocal tract. Accordingly, if speech perception works like visual perception, then recovery of phonetic gestures is not eccentric and does not require eccentric processing by a speech module. It is, instead, yet another instance of recovery of distal events by means of lawfully-generated structure in proximal stimulation.

The general point we hope to make is that, arguably, all perception is heteromorphic, with dimensions of percepts always corresponding to those of distal events, not to distal-event free descriptions of proximal stimuli. Speech is not special in that regard. A more specific point is that even if evidence were to show that speech perceivers do access their speech-motor systems, that perceptual process would not be needed to provide the reason why listeners' percepts are heteromorphic. The reason percepts are heteromorphic is that perceivers universally use proximal stimuli as information about events taking place in the world; they do not use them as perceptual objects *per se*.

Are phonetic gestures public or private?

Although in "Perception of the speech code" and "The motor theory revised," evidence that listeners recover gestures is the only *evidence* cited in favor of the view that perceivers access their speech motor systems, that evidence is not the only *reason* why the motor theorists and other theorists invoke a construct inside the perceiver rather than the proximal stimulation outside to explain why the percept has the character it does. A very important reason why, for the motor theorists, the proximal stimulation is not by itself sufficient to specify phonetic gestures is that, in their view, phonetic gestures are abstract control structures corresponding to the speakers intentions, but not to the movements actually taking place in the vocal tract. If phonetic gestures aren't "out there" in the vocal tract, then they cannot be analogous to other distal events, because they cannot, themselves, lawfully structure the acoustic signal.

In our view, this characterization of phonetic gestures is mistaken, however. We can identify two considerations that appear to support it, but we find neither convincing. One is that any gesture of the vocal tract is merely a token action. Yet perceivers do not just recognize the token, they recognize it *as* a member of a larger linguistically-significant category. That seems to localize the thing perceived in the mind of the perceiver, not in the mouth of the talker. More than that, the same collections of token gestures may be identified as tokens of different categories by speakers of different languages. (So, for example, speakers of English may identify a voiceless unaspirated stop in stressed syllable-initial position as a /b/, whereas speakers of languages in which voiceless unaspirated stops can appear stressed-syllable initially may identify it as an instance of a /p/.) Here, it seems, the information for category membership cannot possibly be in the gestures themselves or in the proximal stimulation; it must be in the head of the perceiver. The second consideration is that coarticulation, by most accounts, prevents nondestructive realization of phonetic gestures in the vocal tract. We briefly address both considerations.

Yet another analogy: There are chairs in the world that do not look very much like prototypical chairs. Recognizing them as chairs may require learning how people typically use them (learning their "proper function" in Millikan's terms [1984]). By most accounts, learning involves some enduring change inside the perceiver. Notice, however, that even if it does, what makes the token chair a chair remains its properties and its use in the world prototypically as a chair. Too, whatever perceivers may learn about that chair and about chairs in general is only what they learn; the chair itself and the means by which its type-hood can be identified remain unquestionably out there in the world. Phonetic gestures and phonetic segments are like chairs (in this respect). Token instances of bilabial closure are members of a type because the tokens all are products of a common coupling among jaw and lips realized in the vocal tract of talkers who achieve bilabial closure. Instances of bilabial closure in stressed-syllable-initial position that have a particular timing relation to a glottal opening gesture are tokens of a phonological category, /b/, in some languages and of a different category, /p/, in others because of the different ways that they are deployed by members of the different language communities. That differential deployment is what allowed descriptive linguists to identify members of phonemic categories as such, and presumably it is also what allows language learners to acquire the phonological categories of their native language. By most accounts, when language learners discover the categories of their language, the learning involves enduring changes inside the learner. However, even if it does, it is no more the case that the phonetic gestures or the phonetic segments move inside the mind than it is that chairs move inside when we learn how to recognize them as such. What we have learned is what we *know* about chairs and phonetic segments; it is not the chairs or the phonetic segments themselves. They remain outside.

Turning to coarticulation, it is described in the motor theory as "encoding," by Ohala (e.g., 1981) as "distortion," by Daniloff and Hammarberg (1973) as "assimilation" and by Hockett (1955) as "smashing" and "rubbing together" of phonetic segments (in the way that raw eggs would be smashed and rubbed together were they sent through a wringer). None of these characterizations is warranted, however. Coarticulation may instead be characterized as gestural layering—a temporally staggered realization of gestures that sometimes do and sometimes do not share one or more articulators.

In fact, this kind of gestural layering occurs commonly in motor behavior. When someone walks, the movement of his or her arm is seen as pendular. However, the surface movement is a complex (layered) vector including not only the

swing of the arm, but also movement of the whole body in the direction of locomotion. This layering is not described as "encoding," "distortion" or even as assimilation of the arm movement to the movement of the body as a whole. And for good reason; that is not what it is. The movement reflects a convergence of forces of movement on a body segment. The forces are separate for the walker, information in proximal stimulation allows their parsing (Johansson, 1973), and perceivers detect their separation.

There is evidence already suggesting that at least some of coarticulation is gestural layering (Carney & Moll, 1971; Öhman, 1966; also see Browman & Goldstein, in press a), not encoding or distortion or assimilation. There is also convincing evidence that perceivers recover separate gestures more-or-less in the way that Johansson suggests they recover separate sources of movement of body segments in perception of locomotion. Listeners use information for a coarticulating segment that is present in the domain of another segment as information for the coarticulating segment itself (e.g., Fowler, 1984; Mann, 1980; Whalen, 1984); they do not hear the coarticulated segment as assimilated or, apparently, as distorted or encoded (Fowler, 1981; 1984; Fowler & Smith, 1986).

Our colleagues Catherine Browman and Louis Goldstein (1985, 1986, in press a,b) have proposed that phonetic primitives of languages are gestural, not abstract featural. Our colleague Elliot Saltzman (1986; Saltzman & Kelso, 1987; see also, Kelso, Saltzman, & Tuller, 1986) is developing a model that implements phonetic gestures as functional couplings among the articulators and that realizes the gestural layering characteristic of coarticulation. To the extent that these approaches both succeed, they will show that phonetic gestures—speakers' intentions—can be realized in the vocal tract nondestructively, and hence can structure acoustic signals directly.

Do listeners need an innate vocal tract synthesizer to recognize acoustic reflections of phonetic gestures? Although it might seem to help, it cannot be necessary, because there is no analogous way to explain how observers recognize most distal events from their optical reflections. Somehow the acoustic and optical reflections of a source must identify the source on their own. In some instances, we begin to understand the means by which acoustic patternings can specify their gestural sources. We consider one such instance next.

How acoustic structure may serve as information for gestures.

We return to the example previously described of listeners' perception of those linguistic dimensions of an utterance that are cued in some way by variation in f_0. A variety of linguistic and paralinguistic properties of an utterance have converging effects on f_0. Yet listeners pull apart those effects in perception.

What guides the listeners' factoring of converging effects of f_0? Presumably, it is the configuration of acoustic products of the several gestures that have effects, among others, on f_0. Intonational peaks are local changes in an f_0 contour that are effected by means that, to a first approximation, only affect f_0; they are produced, largely, by contraction and relaxation of muscles that stretch or shorten the vocal folds (e.g., Ohala, 1978). In contrast, declination is a global change in f_0 that, excepting the initial peak in a sentence, tracks the decline in subglottal pressure (Gelfer et al., 1985; Gelfer et al., 1987). Subglottal pressure affects not only f_0, but amplitude as well, and several researchers have noticed that amplitude declines in parallel with f_0 and resets when f_0 resets at major syntactic boundaries (e.g., Breckenridge, 1977; Maeda, 1976). The parallel decline in amplitude and f_0 constitutes information that pinpoints the mechanism behind the f_0 decline—gradual lung deflation, incompletely offset by expiratory-muscle activity. That mechanism is distinct from the mechanism by which intonational peaks are produced. Evidence that listeners pull apart the two effects on f_0 (Pierrehumbert, 1979; Silverman, 1987) suggests that they are sensitive to the distinct gestural sources of these effects on f_0.

By the same token, f_0 perturbations due to height differences among vowels are not confused by listeners with information for intonational peak height even though f_0 differences due to vowel height are local, like intonational peaks, and are similar in magnitude to differences among intonational peaks in a sentence (Silverman, 1987). The mechanisms for the two effects on f_0 are different, and, apparently, listeners are sensitive to that. Honda (1981) shows a strong correlation between activity of the genioglossus muscle, active in pulling the root of the tongue forward for high vowels, and intrinsic f_0 of vowels. Posterior fibers of the genioglossus muscle insert into the hyoid bone of the larynx. Therefore, contraction of the genioglossus may pull the hyoid

forward, rotating the thyroid cartilage to which the vocal folds attach, and thereby may stretch the vocal folds. Other acoustic consequences of genioglossus contraction, of course, are changes in the resonances of the vocal tract, which reflect movement of the tongue. These changes, along with those in f_0 (and perhaps others as well) pinpoint a phonetic gesture that achieves a vowel-specific change in vocal-tract shape. If listeners can use that configuration of acoustic reflections of tongue-movement (or, more likely, of coordinated tongue and jaw movement) to recover the vocalic gesture, then they can pull effects on f_0 of the vocalic gesture from those for the intonation contour that cooccur with them.

Listeners do just that. In sentence pairs such as "They only feast before fasting" and "They only fast before feasting," with intonational peaks on the "fVst" syllables, listeners require a higher peak on "feast" in the second sentence than on "fast" in the first sentence in order to hear the first peak of each sentence as higher than the second (Silverman, 1987). Compatibly, among steady-state vowels on the same f_0, more open vowels sound higher in pitch than more closed vowels (Stoll, 1984). Intrinsic f_0 of vowels does not contribute to perception of an intonation contour or to perception of pitch. But it is not thrown away by perceivers either. Rather, along with spectral information for vowel height, it serves as information for vowel height (Reinholt Peterson, 1986).

We will not review the literature on listeners' use of f_0 perturbations due to obstruent voicing except to say that it reveals the same picture of the perceiver as the literature on listeners' use of information for vowel height (for a description of the f_0 perturbations: Ohde, 1984; for studies of listeners' use of the perturbations: Abramson & Lisker, 1985; Fujimura, 1971; Haggard, Ambler, & Callow, 1970; Silverman, 1986; for evidence that listeners can detect the perturbations when they are superimposed on intonation contours: Silverman, 1986). As the motor theory and the theory of direct perception both claim, listeners' percepts do not correspond to superficial aspects of the acoustic signal. They correspond to gestures, signaled, we propose, by configurations of acoustic reflections of those gestures.

Does duplex perception reveal a closed speech module?

We return to the phenomenon of duplex perception and consider whether it does convincingly reveal distinct closed and open modules for speech perception and general auditory perception respectively. As noted earlier, duplex perception is obtained, typically, when most of the acoustic structure of a synthetic syllable is presented to one ear, and the remainder—usually a formant transition—is presented to the other ear (refer to Figure 2). In such instances, listeners hear two things. In the ear that gets most of the signal, they hear a coherent syllable, the identity of which is determined by the transition presented to the other ear. At the same time, they hear a distinct, non-speech 'chirp' in the ear receiving the transition. The percept is duplex—the transition is heard as a critical part of a speech syllable, hypothetically as a result of it's being processed by the speech module, and it is heard simultaneously as a non-speech chirp, hypothetically as a result of its being processed also by an open auditory module (Liberman & Mattingly, 1985). Here we offer a different interpretation of the findings.

Whalen and Liberman (1987) have recently shown that duplex perception can occur with monaural or diotic presentation of the base and transition of a syllable. In this case, duplexity is attained by increasing the intensity of the third formant transition relative to the base until listeners hear both an integrated syllable (/da/ or /ga/ depending on the transition) and a non-speech 'whistle' (sinusoids were used for transitions). In the experiment, subjects first were asked to label the isolated sinusoidal transitions as "da" or "ga". Although they were consistent in their labeling, reliably identifying one whistle as "da" and the other as "ga," their overall accuracy was not greater than chance. About half the subjects were consistently right and the remainder were consistently wrong. The whistles are distinct, but they do not sound like "da" or "ga." Next, Whalen and Liberman determined 'duplexity thresholds' for listeners. They presented the base and one of the sinusoids simultaneously and gave listeners control over the intensity of the sinusoid. Listeners adjusted its intensity to the point where they just heard a whistle. At threshold, subjects were able to match these duplex sinusoids to sinusoids presented in isolation. Finally, subjects were asked to identify the integrated speech syllables as "da" or "ga" at sinusoid intensities both 6 dB above and 4 dB below the duplexity threshold. Subjects were consistently good at these tasks yielding accuracy scores well above 90%.

In the absence of any transition, listeners hear only the base and identify it as "da" most of the

time. When a sinusoidal transition is present but at intensities below the duplexity threshold, subjects hear only the unambiguous syllable ("da" or "ga" depending on the transition). Finally, when the intensity of the transition reaches and exceeds the duplexity threshold, subjects hear the "da" or "ga" and they hear a whistle at the same time: i.e., the transition is duplexed.

This experiment reveals two new aspects of the duplex phenomenon. One is that getting a duplex percept requires a sufficiently high intensity of the transition. A second is that the transition integrates with the syllable at intensities below the duplexity threshold. Based on this latter finding, Whalen and Liberman conclude that processing of the sinusoid as speech has priority. It is as if a (neurally-encoded) acoustic signal must first pass through the speech module at which point portions of the signal that specify speech events are peeled off. After the speech module takes its part, any residual is passed on to the auditory module where it is perceived homomorphically. Mattingly and Liberman (1988) refer to this priority of speech processing as "preemptiveness," and Whalen and Liberman (1987) suggest that it reflects the "profound biological significance of speech."

There is another way to look at these findings, however. They suggest that duplex perception does not, in fact, involve the *same* acoustic fragment being perceived in two ways simultaneously. Rather *part* of the transition integrates with the syllable and the *remainder* is heard as a whistle or chirp.[8] As Whalen and Liberman themselves describe it:

> ". . . the phonetic mode takes precedence in processing the transitions, using them for its special linguistic purposes until, having appropriated its share, it passes the *remainder* to be perceived by the nonspeech system as auditory whistles." (Whalen & Liberman, 1987, p. 171; our italics).

This is important, because in earlier reports of duplex perception, it was the apparent perception of the transition in two different ways at once that was considered strong evidence favoring two distinct perceptual systems, one for speech and one for general auditory perception. In addition, research to date has only looked for preemptiveness using speech syllables. Accordingly, it is premature to conclude that speech especially is preemptive. Possibly acoustic fragments integrate preferentially whenever the integrated signal specifies some coherent sound-producing event.

We have recently looked for duplex perception in perception of nonspeech sounds (Fowler and Rosenblum, in press). We predicted that it would be possible to observe duplex perception and preemptiveness whenever two conditions are met: 1) A pair of acoustic fragments is presented that, integrated, specify a natural distal event; and 2) one of the fragments is unnaturally intense. Under these conditions, the integrated event should be preemptive and the intense fragment should be duplexed *regardless* of the type of natural sound-producing event that is involved, whether it is speech or non-speech, and whether it is profoundly biologically significant or biologically trivial.

There have been other attempts to get duplex perception for nonspeech sounds. All the ones of which we are aware have used musical stimuli, however (e.g., Collins, 1985; Pastore, Schmuckler, Rosenblum, & Szczesuil, 1983). We chose not to use musical stimuli because it might be argued that there is a music module. (Music is universal among human cultures, and there is evidence for an anatomical specialization of the brain for music perception (e.g., Shapiro, Grossman, & Gardner, 1981). These considerations led us to choose a non-speech event that evolution could not have anticipated. We chose an event involving a recent human artifact: a slamming metal door.

To generate our stimuli, we recorded a heavy metal door (of a sound-attenuating booth) being slammed shut. A spectrogram of this sound can be seen in Figure 3a. To produce our 'chirp,' we high-pass filtered the signal above 3000 Hz. To produce a 'base,' we low-passed filtered the original signal, also at 3000 Hz (see bottom panels of Figure 3). To us, the high-passed 'chirp' sounded like a can of rice being shaken, while the low-pass-filtered base sounded like a wooden door being slammed shut. (That is, the clanging of the metal door was largely absent.)

We asked sixteen listeners to identify the original metal door, the base and the chirp. The modal identifications of the metal door and the base included mention of a door; however, less than half the subjects reported hearing a door slam. Even so, essentially all of the identifications involved hard collisions of some sort (e.g., boots clomping on stairs, shovel banged on sidewalk). In contrast, no subject identified the chirp as a door sound, and no identifications described hard collisions. Most identifications of the chirp referred to an event involving shaking (tambourine, maracas, castinets, keys).

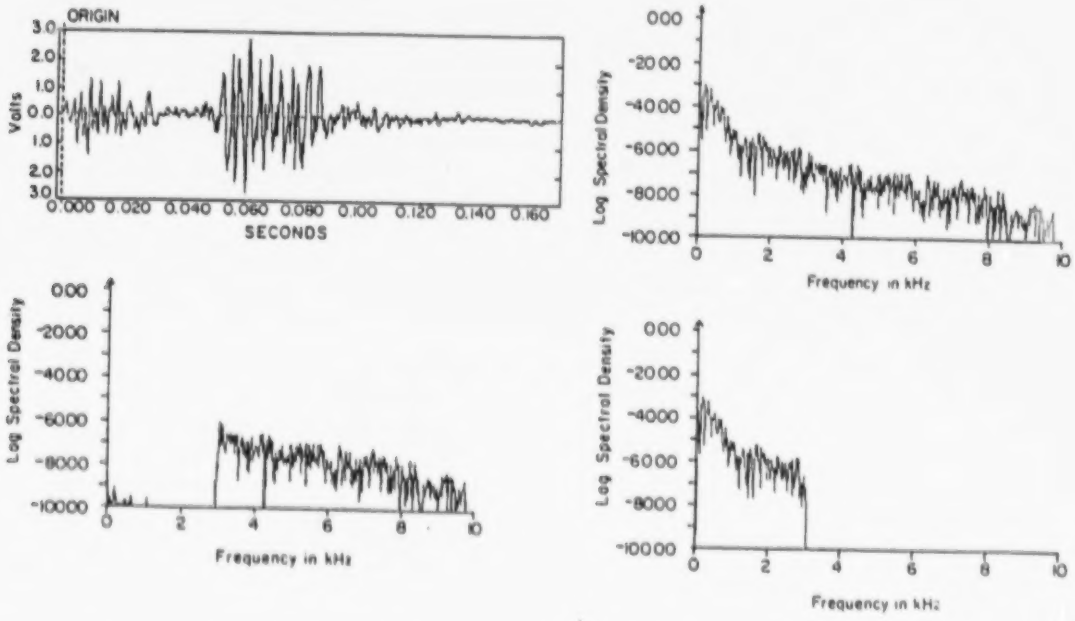


Figure 3. **Display of stimuli used to obtain duplex perception of closing-door sounds.**

Given our metal-door chirp and a high-pass-filtered wooden door slam, subjects could not identify which was a filtered metal door slam and which a filtered wooden door slam. Subjects were consistent in their labeling judgments, identifying one of the chirps as a metal door and the other as a wooden door. However, overall, more of them were consistently wrong than right. On average, they identified the metal-door chirp as the sound of a metal door 31% of the time and the wooden-door chirp as a metal door sound 79% of the time.[9]

To test for duplex perception and preemptiveness, we first trained subjects to identify the unfiltered door sound as a "metal door," the base as a "wooden door" and the upper frequencies of the door as a shaking sound. Next we tested them on stimuli created from the base and the chirp. We created 15 different diotic stimuli. All included the base, and almost all included the metal-door chirp. The stimuli differed in the intensity of the chirp. The chirp was attenuated or amplified by multiplying its digitized voltages by the following values: 0, .05, .1, .15, .2, .9, .95, 1, 1.05, 1.1, 4, 4.5, 5, 5.5 and 6. That is, there were 15 different intensities falling into three ranges; five were well below the natural intensity relationship of the chirp to the base, five were in the range of the natural intensity relation, and five were well above it. Three tokens of each of these stimuli were presented to subjects diotically in a randomized order. Listeners were told that they might hear one of the stimuli, metal door, wooden door or shaking sound, or sometimes two of them simultaneously, on each trial. They were to indicate what they heard on each trial by writing an identifying letter or pair of letters on their answer sheets.

In our analyses, we have grouped responses to the 15 stimuli into three blocks of five. In Figure 4, we have labeled these Intensity Conditions low, medium, and high. Figure 4 presents the results as percentages of responses in the various response categories across the three intensity conditions. We show only the three most interesting (and most frequent) responses. The figure shows that the most frequent response for the low intensity condition is 'wooden door,' the label we asked subjects to use when they heard the base. The most frequent response for the medium condition is 'metal door,' the label we asked subjects to use when they heard the metal door slam. The preferred response for the high-intensity block of stimuli is overwhelmingly 'metal door + chirp,' the response that indicates a duplex percept. The changes in response frequency over the three intensity conditions for each response type are highly significant.

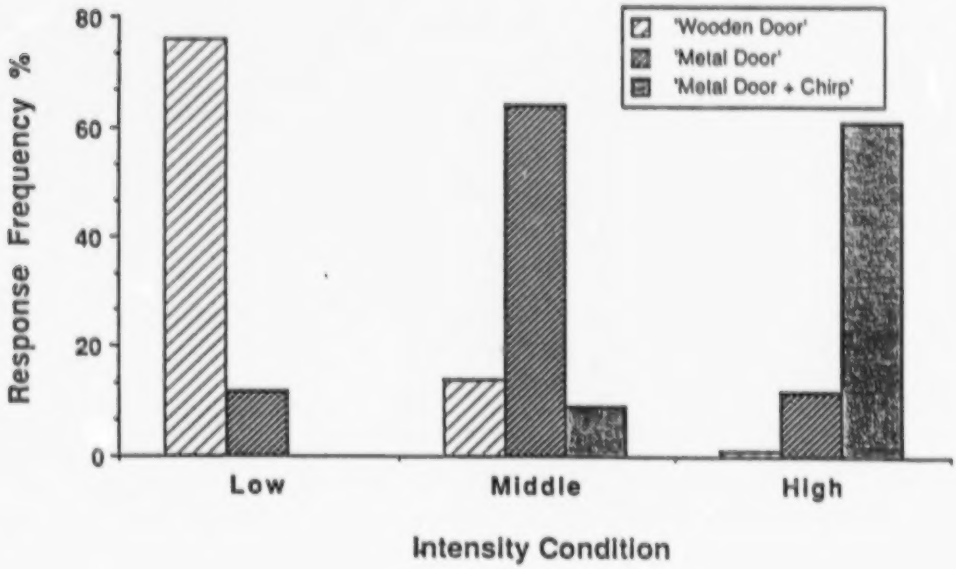


Figure 4. **Percentage of responses falling in the three response categories, "wooden door," "metal door," and "metal door plus shaking sound" across three intensity ranges of the shaking sound.**

Our results can be summarized as follows. First, at very low intensities of the upper frequencies of the door, subjects hear the base only. When the 'chirp' is amplified to an intensity at or near its natural intensity relation to the base, subjects report hearing a metal door the majority of the time. Further amplification of the 'chirp,' leads to reports of the metal door and a separate shaking sound. The percept is duplex, and the metal door slam is preemptive.

There are several additional tests that we must run to determine whether our door slams, in fact, are perceived analogously to speech syllables in procedures revealing duplex perception. If we can show that they are, then we will conclude that an account of our findings that invokes a closed module is inappropriate. Evolution is unlikely to have anticipated metal door slams, and metal-door slams aren't profoundly biologically significant. We suggest alternatively that preemptiveness occurs when a chirp fills a "hole" in a simultaneously presented acoustic signal so that, together the two parts of the signal specify some sound-producing distal event. If anything is left over after the hole is filled, the remainder is heard as separate.

Summary and concluding remarks

We have raised three challenges to the motor theory. We challenge their inference from evidence that phonetic gestures are perceived that speech perception involves access to the talker's own motor system. The basis for our challenge is a claim that dimensions of percepts always conform to those of distal events even in cases where access to an internal synthesizer for the events is unlikely. A second, related, challenge is to the idea that only some percepts are heteromorphic—just those for which we have evolved closed modules. When Liberman and Mattingly write that speech perception is heteromorphic, they mean heteromorphic with respect to structure in proximal stimulation, but they always mean as well that the percept is homomorphic with respect to dimensions of the distal source of the proximal stimulation. We argue that percepts are *generally* heteromorphic with respect to structure in proximal stimulation, but, whether they are or

not, they are always homomorphic with respect to dimensions of distal events. Finally, we challenge the interpretation of duplex perception that ascribes it to simultaneous processing of one part of an acoustic signal by two modules. We suggest, instead, that duplex perception reflects the listener's parsing of acoustic structure into disjoint parts that specify, insofar as the acoustic structure permits, coherent distal events.

Where (in our view) does this leave the motor theory? It is fundamentally right in its claim that listeners perceive phonetic gestures, and also, possibly, in its claim that humans have evolved neural systems specialized for perception and production of phonetic gestures. It is wrong, we believe, specifically in its claims about what those specialized systems do, and generally in the view that closed modules must be invoked to explain why distal events are perceived.

Obviously, we prefer our own, direct-realist, theory, not so much because it handles the data better, but because, in our view, it fits better in a universal theory of perception. But however our theory may be judged in relation to the motor theory, we recognize that we would not have developed it at all in the absence of the important discoveries of the motor theorists that gestures are perceived.

REFERENCES

Abramson, A. & Lisker, L. (1985). Relative power of cues: F_0 shift versus voice timing. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25-32). Orlando, FL: Academic Press.

Breckenridge, J. (1977). *Declination as a phonological process*. Bell Laboratories Technological Memo. Murray Hill, NJ.

Bregman, A. (1987). The meaning of duplex perception. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 95-111). Dordrecht: Martinus Nijhoff.

Browman, C., & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in honor of Peter Ladefoged* (pp. 35-53). Orlando, FL: Academic Press.

Browman, C., & Goldstein, L. (1986). Towards an articulatory phonology. In C. Ewan & J. Anderson (Eds.), *Phonology Yearbook, 3* (pp. 219-254). Cambridge: Cambridge University Press.

Browman, C., & Goldstein, L. (in press a). Tiers in articulatory phonology with some implications for casual speech. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology, I.: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press.

Browman, C., & Goldstein, L. (in press b). Gestural structures and phonological patterns. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception*. Hillsdale: NJ: Lawrence Erlbaum Associates.

Carney, P., & Moll, K. (1971). A cinefluorographic investigation of fricative-consonant vowel coarticulation. *Phonetica, 23*, 193-201.

Collins, S. (1985). Duplex perception with musical stimuli: A further investigation. *Perception and Psychophysics, 38*, 172-177.

Cooper, F., Delattre, P., Liberman, A., Borst, J., & Gerstman, L. (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America, 24*, 597-606.

Daniloff, R., & Hammarberg, R. (1973). On defining coarticulation. *Journal of Phonetics, 1*, 239-248.

Fitch, H., Halwes, T., Erickson, D., & Liberman, A. (1980). Perceptual equivalence of two acoustic cues for stop consonant manner. *Perception and Psychophysics, 27*, 343-350.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Folkins, J., & Abbs, J. (1975). Lip and jaw motor control during speech: Responses to resistive loading of the jaw. *Journal of Speech and Hearing Research, 18*, 207-220.

Folkins, J., & Abbs, J. (1976). Additional observations on responses to resistive loading of the jaw. *Journal of Speech and Hearing Research, 19*, 820-821.

Fowler, C. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research, 46*, 127-139.

Fowler, C. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics, 36*, 359-368.

Fowler, C. (1986a). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*, 3-28.

Fowler, C. (1986b). Reply to commentators. *Journal of Phonetics, 14*, 149-170.

Fowler, C., & Rosenblum, L. D. (in press). Duplex perception: A comparison of monosyllables and slamming of doors. *Journal of Experimental Psychology: Human Perception and Performance*.

Fowler, C., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123-135). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fujimura, O. (1971). Remarks on stop consonants: Synthesis experiments and acoustic cues. In L. Hammerich, R. Jakobson, & E. Zwirner (Eds.), *Form and substance* (pp. 221-232). Copenhagen: Akademisk Forlag.

Gelfer, C., Harris, K., Collier, R., & Baer, T. (1985). Is declination actively controlled? In I. Titze (Ed.), *Vocal-fold physiology: Physiological and biophysics of the voice*. Iowa City: Iowa University Press.

Gelfer, C., Harris, K., & Baer, T. (1987). Controlled variables in sentence intonation. In T. Baer, C. Sasaki, & K. Harris (Eds.), *Laryngeal function in phonation and respiration* (pp. 422-432). Boston: College-Hill Press.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton-Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America, 47*, 613-617.

Hockett, C. (1955). *Manual of phonology*. (Publications in anthropology and linguistics, No. 11). Bloomington: Indiana University Press.

Honda, K. (1981). Relationship between pitch control and vowel articulation. In D. Bless & J. Abbs (Eds.), *Vocal-fold physiology* (pp. 286-297). San Diego: College-Hill Press

Jakobson, R., Fant, C. G. M., & Halle, M. (1951). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

Johansson, G. (1973). Visual perception of biological motion. *Perception and Psychophysics, 14*, 201-211.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. (1984). Functionally-specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative

structures. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 812-832.

Kelso, J. A. S., Saltzman, E., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14,* 29-59.

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology, 15,* 166-171.

Lehiste, I. (1982). Some phonetic characteristics of discourse. *Studia Linguistica, 36,* 117-130.

Lehiste, I., & Peterson, G. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America, 33,* 419-425.

Liberman, A. (1974). The specialization of the language hemisphere (pp. 43-56). In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program.* Cambridge, MA: MIT Press.

Liberman, A. (1982). On finding that speech is special. *American Psychologist, 37,* 148-167.

Liberman, A., Cooper, F., Harris, K., & MacNeilage, P. (1963). A motor theory of speech perception. *Proceedings of the Speech Communication Seminar,* Stockholm: Royal Institute of Technology, D3.

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431-461.

Liberman, A., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic module. *Perception and Psychophysics, 30,* 133-143.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1-36.

Löfqvist, A., Baer, T., McGarr, N. S., & Seider Story, R. (in press). The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America.*

MacDonald, M., & McGurk, H. (1978). Visual influences on speech perception. *Perception and Psychophysics, 24,* 253-257.

Maeda, S. (1976). *A characterization of American English intonation.* Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Mann, V. (1980). Influence of preceding liquid on stop consonant perception. *Perception and Psychophysics, 28,* 407-412.

Mann, V. & Liberman, A. (1983). Some differences between phonetic and auditory modes of perception. *Cognition, 14,* 211-235.

Massaro, D. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mattingly, I. G., & Liberman, A. M. (1988). Specialized perceiving systems for speech and other biologically-significant sounds. In G. Edelman, W. Gall, & W. Cowan (Eds.), *Auditory function: The neurobiological bases of hearing* (pp. 775-793). New York: Wiley.

Mattingly, I. G., & Liberman, A. M. (in press). Speech and other auditory modules. In G. Edelman, W. Gall, & W. Cowan (Eds.), *Signal and sense: Local and global order in perceptual maps.* New York: Wiley.

Millikan, R. (1984). *Language and other biological categories.* Cambridge, MA: MIT Press.

Ohala, J. (1978). Production of tone. In V. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 5-39). New York: Academic Press.

Ohala, J. (1981). The listener as a source of sound change. In C. Masek, R. Hendrick, & M. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178-203). Chicago: Chicago Linguistics Society.

Ohde, R. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *Journal of the Acoustical Society of America, 75,* 224-240.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measures. *Journal of the Acoustical Society of America, 39,* 151-168.

Pastore, R., Schmuckler, M., Rosenblum, L. D., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception and Psychophysics, 33,* 323-332.

Pierrehumbert, J. (1979). The perception of fundamental frequency. *Journal of the Acoustical Society of America, 66,* 363-369.

Porter, R. (1978). Rapid shadowing of syllables: Evidence for symmetry of speech perceptual and motor systems. Paper presented at the meeting of the Psychonomics Society, San Antonio.

Porter, R., & Lubker, J. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research, 23,* 593-602.

Rand, T. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America, 55,* 678-680.

Reed, E. & Jones, R. (1982). *Reasons for realism: Selected essays of James J. Gibson.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Reinholt Peterson, N. (1986). Perceptual compensation for segmentally-conditioned fundamental-frequency perturbations, *Phonetica, 43,* 31-42.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science, 212,* 947-950.

Repp, B. (1987). The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America, 81,* 1100-1109.

Repp, B., Milburn, C. & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. *Perception and Psychophysics, 33,* 333-338.

Rosenblum, L. D. (1987). Towards an ecological alternative to the motor theory of speech perception. *PAW Review* (Technical Report of Center for the Ecological Study of Perception and Action, University of Connecticut), *2,* 25-29.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (pp. 129-144). (Experimental Brain Research Series 15). New York: Springer-Verlag.

Saltzman, E., & Kelso, J. A. S. (1987). Skilled actions: A task-dynamic approach. *Psychological Review, 94,* 84-106.

Schiff, W. (1965). Perception of impending collision. *Psychological Monographs. 79,* No. 604.

Schiff, W., Caviness, J., & Gibson, J. (1962). Persistent fear responses in rhesus monkeys to the optical stimulus of "looming." *Science, 136,* 982-983.

Shapiro, B., Grossman, M., & Gardner, H. (1981). Selective processing deficits in brain-damaged populations. *Neuropsycholgia, 19,* 161-169.

Silverman, K. (1986). F_0 segmental cues depend on intonation: The case of the rise after voiced stops. *Phonetica, 43,* 76-92.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* Unpublished doctoral dissertation, Cambridge University.

Stevens, K. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America, 32,* 47-55.

Stevens, K., & Blumstein, S.(1981). The search for invariant correlates of acoustic features. In P. Eimas & J. Miller (Eds.), *Perspectives on the study of speech* (pp. 1-38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stevens, K., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form* (pp. 88-102). Cambridge, MA: MIT Press.

Stoll, G. (1984). Pitch of vowels: Experimental and theoretical investigation of its dependence on vowel quality. *Speech Communication, 3*, 137-150.

Studdert-Kennedy, M. (1986). Two cheers for direct realism. *Journal of Phonetics, 14*, 99-104.

Studdert-Kennedy, M., & Shankweiler, D. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America, 48*, 579-594.

VanDerVeer, N. (1979). *Ecological acoustics: Human perception of environmental sounds*. Unpublished doctoral dissertation, Cornell University.

Warren, W. & Verbrugge, R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 704-712.

Whalen, D. (1984). Subcategorical mismatches slow phonetic judgments. *Perception and Psychophysics, 35*, 49-64.

Whalen, D. & Liberman, A. (1987). Speech perception takes precedence over nonspeech perception. *Science, 237*, 169-171.

FOOTNOTES

*In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, in press.

†Also Dartmouth College, Hanover, New Hampshire.

††Also University of Connecticut, Storrs. Now at the University of California at Riverside, Department of Psychology.

[1]There is a small qualification to the claim that listeners cannot tell what contributions visible and audible information each have to their perceptual experience in the McGurk effect. Massaro (1987) has shown that effects of the video display can be reduced but not eliminated by instructing subjects to look at, but to ignore, the display.

[2]Liberman et al. identify a cipher as a system in which each unique unit of the message maps onto a unique symbol. In contrast, in a code, the correspondence between message unit and symbol is not 1:1.

[3]Liberman et al. propose to replace the more conventional view of the features of a phoneme (for example, that of Jakobson, Fant & Halle, 1951) with one of features as "implicit instructions to separate and independent parts of the motor machinery" (p. 446).

[4]With one apparent slip on page 2: "The objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands. . . ."

[5]One can certainly challenge the idea that listeners recover the very gestures that occurred to produce a speech signal. Obviously there are no gestures at all responsible for most synthetic speech or for "sine-wave speech" (e.g., Remez, Rubin, Pisoni, & Carrell, 1981) and quite different behaviors underlie a parrot's or mynah bird's mimicking of speech. The claim that we argue is incontrovertible is that listeners recover gestures from speech-like signals, even those generated in some other way. (We direct realists [Fowler, 1986a,b] would also argue that "misperceptions" (hearing phonetic gestures where there are none) can only occur in limited varieties of ways—the most notable being signals produced by certain mirage-producing human artifacts, such as speech synthesizers or mirage-producing birds. Another, however, possibly, includes signals produced to mimic those of normal speakers by speakers with pathologies of the vocal tract that prevent normal realization of gestures.)

[6]There are two almost orthogonal perspectives from which perception can be studied. On the one hand, investigators can focus on processes inside the perceiver that take place from the time that a sense organ is stimulated until a percept is achieved or a response is made to the input. On the other hand, they can look outside the perceiver and ask what, in the environment, the organism under study perceives, what information in stimulation to the sense organs allows perception of the things perceived, and finally, whether the organisms in fact use the postulated information. Here we focus on this latter perspective, most closely associated with the work of James Gibson (e.g., 1966; 1979; Reed & Jones, 1982).

[7]It is easy to find examples in which perception is heteromorphic with respect to the proximal stimulation and homomorphic with respect to distal events—looming, for example. We can also think of some examples in which perception appears homomorphic with respect to proximal stimulation, but in the examples we have come up with, they are homomorphic with respect to the distal event as well (perception of a line drawn by a pencil, for example), and so there is no way to decide whether perception is of the proximal stimulation or of the distal event. We challenge the motor theorists to provide an example in which perception is homomorphic with structure in proximal stimulation that is not also homomorphic with distal event structure. These would provide convincing cases of proximal stimulation perception.

[8]Bregman (1987) considers duplex perception to disconfirm his "rule of disjoint allocation" in acoustic scene analysis by listeners. According to the rule, each acoustic fragment is assigned in perception to one and only one environmental source. It seems, however, that duplex perception does not disconfirm the rule.

[9]Using a more sensitive, AXB, test, however, we have found that listeners can match the metal door chirp, rather than a wooden door chirp, to the metal door slam at performance levels considerably better than chance.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 118-136

Competence and Performance in Child Language*

Stephen Crain† and Janet Dean Fodor††

1. INTRODUCTION

This paper presents the results of our recent experimental investigations of a central issue in linguistic theory: which properties of human language are innately determined? There are two main sources of information to be tapped to find the answer to this question. First, universal properties of human languages are plausibly (even if not necessarily) taken to be innately determined. In addition, properties that emerge in children's language in the absence of decisive evidence in their linguistic input are reasonably held to be innate. Clearly, it would be most satisfactory if these two diagnostics for what is innate agreed with each other. In some cases they do. For example, there is a universal principle favoring transformational movement of phrases rather than of lexical categories e.g., topicalization of noun phrases but not of nouns. To the best of our knowledge children abide by this principle; they may hear sentences such as *Candy, you can't have now,* but they don't infer that nouns can be topicalized. If they did, they would say things like **Vegetables, I won't eat the.* But this is not an error characteristic of children. Instead, from the moment they produce topicalized constructions at all, they apparently produce correct NP-topicalized forms such as *The vegetables, I won't eat.*

In recent years, this happy convergence of results from research on universals and research on acquisition has been challenged by experimental studies reporting various syntactic failures on the part of children. The children in these experiments are apparently violating putatively universal phrase structure principles or constraints on transformations. Failure to demonstrate early knowledge of syntactic principles is reported by Jakubowicz (1984), Lust (1981), Matthei (1981, 1982), Phinney (1981), Roeper (1986), Solan and Roeper (1978), Tavakolian (1978, 1981) and Wexler and Chien (1985). Some explanation is clearly called for if a syntactic principle is respected in all adult languages but is not respected in the language of children.

Assuming that the experimental data accurately reflect children's linguistic competence, there are several possible responses to the unaccommodating data. The most extreme would be to give up the innateness claim for the principle in question. One might look for further linguistic data which show that it isn't universal. Or one might abandon the hypothesis that all universal principles are innate. For instance, Matthei (1981) obtained results that he interpreted as evidence that universal constraints on children's interpretation of reciprocals are learned, not innate. However, this approach is plausible only if one can offer some other explanation (e.g., functional explanation) for why the constraints should be universal. But this is not always easy; as Chomsky (1986) has emphasized, many properties of natural language are arbitrary and have no practical motivation.

A different response to the apparent failure of children to respect constraints believed to be innate is to argue that the constraints are as yet inapplicable to their sentences. The claim is that as soon as a child's linguistic analyses have reached the level of sophistication at which a

This research was supported in part by NSF Grant BNS 84-18537, and by a Program Project Grant to Haskins Laboratories from the National Institute of Child Health and Human Development (HD-01994). The studies reported in this paper were conducted in collaboration with several friends and colleagues: Henry Hamburger, Paul Gorrell, Howard Lasnik, Cecile McKee, Keiko Murasugi, Mineharu Nakayama, Jaya Sarma and Rosalind Thornton. We thank them for their permission to gather this work together here.

universal constraint becomes relevant, then that constraint will be respected. For example, Otsu (1981) has argued that children who give the appearance of violating a universal constraint on extraction may not yet have mastered the structure to which the constraint applies; they may have only some simpler approximation to the construction, lacking the crucial property that engages the constraint. We will discuss the evidence for this below.

A different approach also accepts the recalcitrant data as valid, but rejects the inference that the data are inconsistent with the innateness hypothesis. It is pointed out that it is possible for a linguistic principle to be innately encoded in the human brain and yet not accessible to the language faculty of children at early stages of language acquisition. The principle in question might be biologically timed to become effective at a certain maturational stage. Like aspects of body development (e.g., the secondary sex characteristics), linguistic principles might lie dormant for many years. One recent proposal invoking linguistic maturation, by Borer and Wexler (1987), contends that a syntactic principle underlying verbal passives undergoes maturational development. They maintain that before a critical stage of maturation is reached, children are unable to produce or comprehend passive sentences (full verbal passives with by-phrases).

It may eventually turn out that the innateness hypothesis must be augmented by maturation assumptions in certain cases. But such assumptions introduce new degrees of freedom into the theory, so its empirical claims are weakened. Unless some motivated predictions can be made about exactly **when** latent knowledge will become effective, a maturational approach is compatible with a much wider range of data than the simplest and strongest version of the innateness hypothesis, viz. that children have access to the same set of universal principles at all stages of language development. This more restricted position is the one to be adopted until or unless there is clear evidence to the contrary, e.g., clear evidence of a period or a stage at which all children violate a certain constraint, in all constructions to which it is applicable, simple as well as complex, and in all languages. So far, no such case has been demonstrated.[1]

Our research has taken a different approach. We argue that the experimental data do **not** unequivocally demonstrate a lack of linguistic knowledge. We do not deny that children do sometimes misinterpret sentences. But the proper interpretation of such failures is complicated by the existence of a variety of potentially confounding factors. Normal sentence comprehension involves lexical, syntactic, semantic, pragmatic and inferential abilities, and the failure of any one of these may be responsible for poor performance on an experimental task. It is crucial, therefore, to develop empirical methods which will distinguish between these various factors, so that we can determine exactly where a child's deficiencies lie. Until this has been done, one cannot infer from children's imperfect performance that they are ignorant of the grammar of their target language.

In fact it can be argued in many cases that it is non-syntactic demands of the task which are the cause of children's errors. We propose that task performance is weak at first and improves with age in large part because of maturation of non-linguistic capacities such as short term memory or computational ability, which are essential in the efficient practical application of linguistic knowledge. This does not deny that many aspects of language must be learned and that there is a time when a child has not yet learned them. But our interpretation of the data does make it plausible that young children know more of the adult grammar than has previously been demonstrated, and also, most significantly, that their early grammars do abide by universal principles.

In support of this non-linguistic maturation hypothesis, we have reexamined and supplemented a number of earlier experimental findings with demonstrations that nonsyntactic factors were responsible for many of the children's errors. The errors disappear or are greatly reduced when these confounding factors are suitably controlled for. In this paper, we report on a series of experimental studies along these lines concerned with three kinds of nonsyntactic factors in language performance: parsing, plans, and presuppositions. We will argue that these other factors are crucially involved in the experimental tasks by which children demonstrate their knowledge, and that they impose significant demands on children. If we underestimate the demands of any of these other components of the total task we thereby underestimate the extent of the child's knowledge of syntax. As a result, the current estimate of what children know about language is misleading.

2. CHILDREN'S ERRORS IN COMPREHENSION

In this section we attempt to identify and isolate several components of language-related skills, in order to gain a better understanding of each, and to clarify the relationship between the innateness hypothesis and early linguistic knowledge. Very little work has been done on this topic. The majority of language development studies seem to take it for granted that the experimental paradigms provide a direct tap into the child's linguistic competence. An important exception is a study by Goodluck and Tavakolian (1982), in which improved performance on a relative clause comprehension task resulted from simplification of other aspects of the syntax and semantics of the stimulus sentences (i.e., the use of intransitive rather than transitive relative clauses, and relative clauses with one animate and one inanimate noun phrase rather than two animates). The success of these manipulations is exactly in accord with our general hypothesis about the relation between competence and performance. As other demands on the child's performance are reduced, greater competence is revealed.

Our experiments focus on three factors involved in many child language experiments which may interfere with estimation of the extent of children's linguistic knowledge in tasks which are designed to measure sentence comprehension. These factors—parsing, presupposition and plans—are of interest in their own right, but have received very little attention in previous research on syntax acquisition. In this section we will review our recent work on these topics. In the following section we will turn to an alternative research strategy for assessing children's knowledge, the technique of elicited production.

Parsing. Sentence parsing is a complex task which is known to be governed (in adults) by various decision strategies that favor one structural analysis over another where both are compatible with the input word sequence. Even adults make parsing errors, and it would hardly be surprising, given the limited memory and attention spans of children, to discover that they do too. These parsing preferences must somehow be neutralized or factored out of an experimental task whose objective is the assessment of children's knowledge of syntactic rules and constraints.

Plans. The formation of an action plan is an important aspect of any comprehension task involving the manipulation of toys or other objects. If the plan for manipulating objects appropriately in response to a test sentence is necessarily complex to formulate or to execute, its difficulty for a child subject may mask his correct comprehension of the sentence. Thus we need to develop a better understanding of the nature and relative complexity of such plans, and also to devise experimental paradigms in which their impact on performance is minimized.

Presuppositions. A variety of pragmatic considerations must also be taken into account, such as the contextual fixing of deictic reference, obedience to cooperative principles of conversation, and so forth. In particular, our research suggests that test sentences whose pragmatic presuppositions are unsatisfied in the experimental situation are also unlikely to provide results allowing an accurate assessment of a child's knowledge of syntactic principles. It is necessary to establish which kinds of presuppositions children are sensitive to, and to ensure that these are satisfied in experimental tasks.

2.1. Parsing

2.1.1. Subjacency. One universal constraint which should be innate is Subjacency. Subjacency prohibits extraction of constituents from various constructions, including relative clauses. However, in an experimental study by Otsu (1981), many children responded as if they allowed extraction from relative clauses in answering questions about the content of pictures. For example, children saw a picture of a girl using a crayon to draw a monkey who was drinking milk through a straw. They were then asked to respond to question (1).

(1) What is Mary drawing a picture of a monkey that is drinking milk with?

Otsu found that many children responded to (1) in a way that appeared to violate Subjacency. In this case, the answer that is in apparent violation of Subjacency is "a straw." This is because "a straw" is appropriate only if the *what* has been moved from a position in the *monkey drinking milk* clause as shown in (2a), rather than from the *Mary drawing picture* clause as shown in (2b).

(2) a) *What is Mary drawing a picture of a monkey [that is drinking milk with _]?

b) What is Mary drawing a picture of a monkey [that is drinking milk] with _?

But the *monkey drinking milk* clause is a relative clause, and Subjacency prohibits the *what* from moving out of it. Thus the only acceptable structure is (2b), and the only acceptable answer is "a crayon." If these data are interpreted solely in terms of children's grammatical **knowledge,** then the conclusion would then have to be that knowledge of Subjacency sets in quite late in at least some children.

As we noted earlier, Otsu suggested that the innateness of Subjacency could be salvaged by showing that the children who appeared to violate Subjacency had not yet mastered the phrase structure of relative clauses (of sufficient complexity to contain an extractable noun phrase). When he conducted an independent test of knowledge of relative clause structure, he found, as predicted, a correlation between phrase structure and Subjacency application in the children's performance. However, the children's performance was still surprisingly poor: 25% of the children who were deemed to have mastered relative clauses gave responses involving ungrammatical Subjacency violating extractions from relative clauses.

We have argued (Crain & Fodor, 1984) for an alternative analysis of Otsu's data, which makes it possible to credit children with knowledge of **both** phrase structure principles **and** constraints on transformations from an early age. We claim that children's parsing routines can influence their performance on the kind of sentences used in the Subjacency test; in particular, that there are strong parsing pressures encouraging subjects to compute the ungrammatical analysis of such sentences. Until a child develops sufficient capacity to override these parsing pressures, they may mask his syntactic competence, making him look as if he were ignorant of Subjacency.

A powerful general tendency in sentence parsing by adults is to attach an incoming phrase low in the phrase marker if possible. This has been called Right Association; see Kimball (1973), Frazier and Fodor (1978). In sentence (3), for example, the preferred analysis has *with NP* modifying *drinking milk* rather than modifying *drawing a picture,* even though in this case both analyses are grammatically well-formed because there has been no WH-movement.

(3) Mary is drawing a picture of a monkey that is drinking milk with NP.

To see how strong this parsing pressure is, note how difficult it is to get the sensible interpretation of (3) when *a crayon* is substituted for NP. This Right Association preference is still present if the NP in (3) is extracted, as in (1). The word *with* in (1) still coheres strongly with the relative clause, rather than with the main clause. The result is that the analysis of (2) that is most immediately apparent is the ungrammatical (2a) in which *what* has been extracted from the relative clause. Since this 'garden path' analysis is apparent to most adults, it is hardly surprising if some of Otsu's child subjects were also tempted by it and responded to (1) in the picture verification task by saying "a straw" rather than "a crayon."

We conducted several experiments designed to establish the plausibility of this claim that the relatively poor performance of children on sentences like (1) is due to parsing pressures rather than to ignorance of universal constraints. In the first experiment, we tested children and adults on complement-clause questions as in (4). Subjacency does not prohibit extraction from complement clauses, so if there were no Right Association effects this sentence should be fully ambiguous, with both interpretations equally available.

(4) What is Bozo watching the dog jump through?

That is, given a picture in which Bozo the clown is looking through a keyhole at a dog jumping through a hoop, it would be correct to say either the "the keyhole" or "a hoop." Intuitively, though, the interpretations are highly skewed for adults, with a strong preference for the Right Association interpretation ("hoop") in which the preposition attaches within the lower clause. Our experiment showed that the same is true for children. We tested twenty 3- to 5-year-olds (mean age 4;6) on these sentences using a picture verification task just like Otsu's, and 90% of their responses were in accord with the Right Association interpretation.[2]

Thus children and adults alike are strongly swayed by Right Association. This is an important result. To the best of our knowledge the question of whether children's parsing strategies resemble those of adults has not previously been investigated. But children certainly should show the same preferences as adults, if the human sentence parsing mechanism is innately structured. And the parsing mechanism certainly should be innately structured, because it would be pointless to be born knowing a lot of facts about language if one weren't also born knowing how to use those facts for speaking and understanding. It is satisfying, then, to have shown that children

exhibit Right Association. And the fact that they do offers a plausible explanation for why so many of them failed Otsu's Subjacency test—they were listening to their parsers rather than to their grammars.

Our other experiments in support of this conclusion were designed to show that even people whose knowledge of Subjacency is not in doubt—i.e., adults—are also tempted to violate Subjacency when it is in competition with Right Association. We ran Otsu's Subjacency test on adults just as he did with the children. The adults gave Subjacency-violating low attachment responses to 21% of these questions. This was not quite as high a rate as for the children, but as we have noted, adults surely have a greater capacity than children do for checking an illicit analysis and shifting to a less preferred but well-formed analysis before they commit themselves to a response. In an attempt to equalize adult self-monitoring capacities with those of children, we re-ran the Subjacency experiment with an additional distracting task (= listening for a designated phoneme in the stimulus sentence). Under these conditions the adults gave Subjacency violating responses to 29% of the relative clause constructions, a slightly higher rate than the 25% for Otsu's child subjects.

Escalating still further, we changed the sentences so that the grammatically well-formed analysis was semantically or pragmatically anomalous, as in (5).

(5) What color hat is Barbara drawing a picture of an artist with?

Under these circumstances, where the semantics clearly favored the Subjacency-violating analysis, 75% of adults' responses violated Subjacency. This makes it very clear that linguistic competence may not always be revealed by linguistic performance.

Finally, we ran another study, in which we asked adults to classify sentences as ambiguous or unambiguous. The sentences were spoken in turn with only a few seconds between them, and there were 72 of them, so the task was fairly demanding. The materials included complement questions like (4) and relative clause questions like (1), as well as ambiguous and unambiguous control sentences of many varieties. The results showed a 62% ambiguity detection rate for the ambiguous control sentences, with a 16% 'false alarm' rate for the unambiguous control sentences. Thus the subjects were able to cope with the task tolerably well, though not perfectly. What was interesting was that the ambiguity of the complement questions was detected only 48% of the time, in line with our claim that Right Association obscures the alternative reading with the prepositional phrase in the main clause. And most interesting of all was that 80% of the relative clause questions were judged to be ambiguous, even though Subjacency prohibits one analysis and renders them unambiguous. Our explanation for this extraordinary result is that the subjects first computed the Right Association analysis favored by their parsing routines, then recognized that this was unacceptable because of Subjacency, and so rejected it in favor of the analysis with the prepositional phrase in the main clause. We assume that it was this rapid shift from one analysis to the other that gave our subjects such a strong impression that these sentences were ambiguous. Note that if this misanalysis-with-revision occurs 80% of the time for adults, only a slight handicap in children's ability to revise would be sufficient to account for their errors.

To sum up: we still have no positive proof that Subjacency is innate, but at least now there is no evidence against it. Our experiments make it plausible that children as young as can be tested are like adults **both** with respect to their knowledge of this universal constraint **and** with respect to their parsing routines—they are just not very good yet at coping with conflicts between the two.

2.1.2. Backward pronominalization. A fundamental constraint on natural language is the structure dependence of linguistic rules. The innateness hypothesis implies that children's earliest grammars should also exhibit structure dependence—even if their linguistic experience happens to be equally compatible with structure-independent hypotheses. However, it has been proposed that children initially hypothesize a structure-independent constraint on anaphora, prohibiting all cases of backward pronominalization (Solan, 1983).[3] Backward pronominalization consists of coreference between a noun phrase and a preceding pronoun, as indicated by the indices in (6).

(6) That he_i kissed the lion made the $duck_i$ happy.

We will argue that children do in fact permit backward pronominalization, subject to structure-dependent constraints. We contend that the appearance of a general restriction against backward pronominalization is due to a parsing preference for the alternative 'extra-sentential'

reading of the pronoun in certain comprehension tasks. The results of a new comprehension methodology show that children as young as 2;10 admit the same range of interpretations for pronouns as adults do.

Two sources of evidence have been cited as evidence that children up to 5 or 6 years uniformly reject backward pronominalization. First, children who are asked to repeat back a sentence such as (7) often respond by converting it into a forward pronominalization construction, as in (8) (Lust, 1981).

(7) Because she was tired, Mommy was sleeping.

(8) Because Mommy was tired, she was sleeping.

The fact that these children took the trouble to exchange the pronoun and its antecedent certainly indicates that they disfavor backward pronominalization in their own productions. But it does not show that the backward pronominalization interpretation is not compatible with the child's **grammar,** as suggested by Solan (1983). To the contrary, the conversion of (7) to (8) shows that children do accept backward pronominalization in comprehension; for they would think of (8) as an acceptable variant of (7) only if they were interpreting the pronoun in (7) as coreferential with the subsequent lexical noun phrase (Lasnik & Crain, 1985).

Second, it has been found that when the acting-out situation for a sentence like (6) includes a potential referent for *he* other than the duck (e.g., a farmer), this unmentioned object is usually favored by the children as the referent of the pronoun (Solan, 1983; Tavakolian, 1978). In contrast to the prevailing view, we would attribute this to a parsing preference for the extra-sentential interpretation of the pronoun; it does not have to be taken as evidence that children have a grammatical prohibition against backward anaphora. Our suggestion, then, is that children's **knowledge** might be comparable to that of adults, even if their performance differs.

It is particularly important to keep this distinction in mind for potentially ambiguous constructions such as these. When a sentence has more than one possible interpretation, the interpretation that children select can tell us which interpretation they prefer; it **cannot** show that others are unavailable to them. After all, adults also exhibit biases in connection with ambiguous constructions, but this does not lead us to accuse them of ignorance of alternative interpretations. To establish how much children actually do know, we should look for the factors that might be biassing their interpretations, and also for ways of minimizing this bias so that interpretations which are less preferred but nevertheless acceptable to them have a chance of showing through.

The most likely general source of bias against backward pronominalization is the fact that interpretation of the pronoun would have to be delayed until the antecedent is encountered later in the sentence. This retention of uninterpreted items may strain a child's limited working memory. There is some evidence for this speculation. Hamburger and Crain (1984) have noted that children show a tendency to interpret adjectives immediately, without waiting for the remainder of the noun phrase, even in cases where this leads them to give incorrect responses. And Clark (1971) has observed errors attributable to children's tendency to act out a clause immediately without waiting for other clauses in the sentence. The only way to interpret the pronoun immediately in a sentence like (6) is to assign it an extra-sentential referent, as children typically do.

If this proposal is correct, it should be that children will accept backward pronominalization in an experimental task that presses subjects to access **every** interpretation they can assign to a sentence. Crain and McKee (1985) used a true/false paradigm in which subjects judge the truth value of sentences against situations acted out by the experimenter. The sentences were as in (9), where either a coreferential reading or an extra-sentential reading of the pronoun is possible.

(9) When he went into the barn, the fox stole the food.

On each trial, a child heard a sentence following a staged event acted out by one of two experimenters, using toy figures and props. The second experimenter manipulated a puppet, Kermit the Frog. Following each event, Kermit said what he thought had happened on that trial. The child's task was to indicate whether or not the sentence uttered by Kermit accurately described what had happened. Children were asked to feed Kermit a cookie if he said the right thing, that is, if what he said was what really happened. In this way, 'true' responses were encouraged in the experimental situation. But sometimes Kermit would say the wrong thing, if he wasn't paying close attention. When this happened, the child was asked to make Kermit eat a rag. (In pilot

work without the rag ploy, we had found that children were reluctant to say that Kermit had said something wrong.)

To test for the availability of both interpretations of an ambiguous sentence like (9), children judged it twice during the course of the experiment, once following a situation in which a fox stole some chickens from inside a barn (for the backward pronominalization interpretation), and once following a situation in which a man stole some chickens while a fox was in a barn (for the extra-sentential interpretation).

Children accepted the backward anaphora reading for all the ambiguous sentences 73 of the time. The extra-sentential reading was accepted 81 of the time, but the difference was not significant. Much the same results were obtained even for the 7 youngest children, whose ages were from 2;10 to 3;4. Only two of the 62 subjects consistently rejected the backward anaphora reading. Thus most children find the backward anaphora reading acceptable, although it might not be preferred if they were forced to choose between interpretations, as in previous comprehension studies.

We should note that a variety of control sentences were also tested to rule out other, less interesting, explanations of the children's performance. For example, the children rejected sentence (10) following a situation in which Strawberry Shortcake did eat an ice cream, but not while she was outside playing. This shows that they were not simply ignoring the subordinated clauses of sentences in deciding whether to accept or reject them.

(10) When she was outside playing, Strawberry Shortcake ate an ice cream.

Sentences like (11) were also tested in order to establish that subjects were not merely giving positive responses to all sentences, regardless of their grammatical properties.

(11) He stole the food when the fox went into the barn.

The difference between (11) and the acceptable backward pronominalization in (9) is that in (11) the pronoun is in the higher clause and c-commands *the fox*, while in (9) the pronoun is in the subordinate clause and does not c-command *the fox*. (A node A in a phrase marker is said to c-command a node B if there is a route from A to B which goes up to the first branching node above A, and then down to B. Note that c-command is a structure-dependent relation.) There is a universal constraint that prohibits a pronoun from c-commanding its antecedent. And indeed the children did reject (11) 87% of the time. Note that this positive result shows that the children have early knowledge not only of the absence of linear sequence conditions on pronominalization, but also of the existence of structural conditions such as c-command. (See also Lust, 1981, and Goodluck, 1986.)

2.1.3. Subject/Auxiliary Inversion. Another study (Crain & Nakayama, 1987) also explored the tie between children's errors in acquisition tasks and sentence processing problems. This study was designed to test whether children give structure-dependent or structure-independent responses when they are required to transform sentences by performing Subject/Auxiliary inversion. As Chomsky (1971) pointed out, transformational rules are universally sensitive to the structural configurations in the sentences to which they apply, not just to the linear sequence of words.

The procedure in this study was simply for the experimenter to preface declaratives like (12) with the carrier phrase "Ask Jabba if ...," as in (13).

(12) The man who is running is bald.

(13) Ask Jabba if the man who is running is bald.

The child then had to pose the appropriate yes/no questions to Jabba the Hutt, a figure from "Star Wars" who was being manipulated by one of the experimenters. Following each question, Jabba was shown a picture and would respond "yes" or "no."

The sentences all contained a relative clause modifying the subject noun phrase. The correct structure-dependent transformation moves the first verb of the main clause to the front of the sentence, past the whole subject noun phrase, as in (14). An incorrect, structure-independent transformation would be as in (15), where the linearly first verb in the word string (which happens to be the verb of the relative clause) has been fronted.

(14) Is the man who is running bald?

(15) *Is the man who running is bald?

For simple sentences with only one clause such as (16), which are more frequent in a young child's input, both versions of the transformation rule give the correct result.

(16) Ask Jabba if the man is bald.

(17) Is the man bald?

It is only on the more complex sentences that the form of the child's rule is revealed.

The outcome was as predicted by the innateness hypothesis: children never produced an incorrect sentence like (15). Thus, a structure-independent strategy was not adopted in spite of its simplicity and in spite of the fact that it produces the correct question forms in many instances. The findings of this study thus lend further support to the view that the initial state of the human language faculty contains structure-dependence as an inherent property.

The children did make some errors in this experiment, and we observed that most of them were in sentences with a long subject noun phrase and a short main verb phrase, as in (18).

(18) Is the boy who is holding the plate crying ?

By contrast, there were significantly fewer errors in sentences like (19), which has a shorter subject noun phrase and a longer verb phrase.

(19) Is the boy who is unhappy watching Mickey Mouse ?

This kind of contrast is familiar in parsing studies with adults. In particular, Frazier and Fodor (1978) showed that a sequence consisting of a long constituent followed by a short constituent is especially troublesome for the (adult) parsing routines;[4] a short constituent before a long one is much easier to parse. The distribution of the children's errors in the Subject-Auxiliary Inversion task may therefore be indicative not of inadequate knowledge of the inversion rule, but of an adult-like processing sensitivity to interactions between structure and constituent length.

A follow-up study to test this possibility was conducted by Nakayama (1987). Nakayama systematically varied both the length and the syntactic structure of the sentences to be transformed by the children. The children made significantly fewer Subject-Auxiliary Inversion errors in response to embedded questions with short relative clauses (containing intransitive verbs) as compared to those with long relative clauses (containing transitive verbs). With length held constant, the children had more difficulty with relative clauses that had object gaps, as in (20), than with relative clauses that had subject gaps, as in (21) (although this effect was not quite significant).

(20) The ball the girl kicked is rolling.

(21) The boy who was slapped is crying.

The ease of subject gap constructions, as compared to object gap constructions, has been found in a number of other studies in language development, in language impaired populations, and in experiments on adult sentence processing (where the question of syntactic competence is not in doubt). It seems reasonable to interpret these results as confirming that children's error rates in language tests are highly sensitive to the complexity of the sentence parsing that is required.

2.2. Presupposition

Syntactic parsing is not the only factor that has been found to mask knowledge of syntactic principles. Test sentences whose pragmatic presuppositions are unsatisfied in the experimental situation have been found to result in inaccurate assessments of children's structural knowledge. In this section we consider two experiments that point to the relevance of presuppositional content in sentence understanding.

The structures we discuss here are relative clauses and temporal adverbial clauses. A word of clarification is needed before we proceed. Up till now we have restricted the scope of the innate hypothesis to universal constraints (like Subjacency, and structure-dependence), which could not **in principle** be learned from normal linguistic experience (i.e., without extensive corrective feedback). But now we want to extend the innateness hypothesis to a broader class of linguistic knowledge, knowledge of universal types of sentence construction. We cannot plausibly claim that every aspect of these constructions is innate. Rather, every construction will have some aspects that are determined by innate principles, and other aspects that must be learned. And the balance between these two elements varies from construction to construction. So it is perfectly acceptable on theoretical grounds that some constructions should be acquired later than others. However, the innateness hypothesis is not compatible with just **any** order of acquisition. It predicts early acquisition of constructions that Chomsky calls 'core' language, i.e., the constructions that have strong assistance from innate principles with just a few parameters to be set by learners on the basis of experience. It would be surprising to discover that knowledge of **these** constructions was significantly delayed once the relevant lexical items had been learned. In the absence of a plausible explanation, this would put the innateness hypothesis at risk.

We noted in section 1 a range of possible explanations of apparently delayed knowledge of linguistic facts. In the present case they would include the following:

- the construction does not, after all, belong to the core but is 'peripheral' and hence should be acquired late;
- children don't hear this construction until quite late in the course of language development and so could not be expected to know it exists;
- the core principles in question undergo maturation and so are not accessible at early stages of acquisition;
- the experimental data are faulty and children do indeed have knowledge of this construction.

We will argue for this last alternative. And just as in the previously described studies of innate constraints, we will lay the blame for the misleading experimental data on the fact that traditional experimental paradigms do not make sufficient allowance for the limited memory and computational capacities of young children. Once again, our story is that non-linguistic immaturity can create the illusion of linguistic immaturity.

2.2.1. Relative Clauses. Children typically make more errors in understanding sentences containing relative clauses (as in 22) than sentences containing conjoined clauses (as in 23), when comprehension is assessed by a figure manipulation (act-out) task.

(22) The dog pushed the sheep that jumped over the fence.

(23) The dog pushed the sheep and jumped over the fence.

The usual finding that (22) is more difficult for children than (23) up to age 6 years or so has been interpreted as an instance of late emergence of the rules for subordinate syntax in language development (e.g., Tavakolian, 1981). However, though coordination may be innately favored over subordination, it is also true that subordination is ubiquitous in natural language; relative clause constructions are very close to the 'core.' So ignorance of relative clauses until age 6 would stretch the innateness hypothesis.

Fortunately this is not how things stand. Hamburger and Crain (1982) showed that the source of children's performance errors on this task is **not** a lack of syntactic knowledge. By constructing pragmatic contexts in which the presuppositions of restrictive relative clauses were satisfied, they were able to demonstrate mastery of relative clause structure by children as young as 3 years. There are two presuppositions in (22): (i) that there are at least two sheep in the context, and (ii) that one (but only one) of the sheep jumped over a fence prior to the utterance. The reason why previous studies failed to demonstrate early knowledge of relative clause constructions, we believe, is that they did not pay scrupulous attention to these pragmatic presuppositions. For example, subjects were required to act out the meaning of a sentence such as (22) in contexts in which only one sheep was present. The poor performance by young children in these experiments was attributed to their ignorance of the linguistic properties of relative clause constructions. But suppose that a child did know the linguistic properties, but that he also was aware of the associated presuppositions. Such a child might very well be unable to relate his correct understanding of the sentence to the inappropriate circumstances provided by the experiment. Adult subjects may be able to 'see through' the unnaturalness of an experimental task to the intentions of the experimenter, but it is not realistic to expect this of young children.

Following this line of reasoning, Hamburger and Crain (1982) made the apparently minor change of adding two more sheep to the acting out situation for sentence (22), and obtained a much higher percentage of correct responses. The most frequent remaining 'error' was failure to act out the event described by the relative clause, but since felicitous usage presupposes that this event has already occurred, this is not really an error but is precisely the kind of response that is compatible with perfect comprehension of the sentence. This interpretation of the data is supported by the fact that there was a positive correlation between incidence of this response type and age.[5]

We have conducted another series of studies on relative clauses, trying several other techniques for assessing grammatical competence. In one study, we employed a picture verification paradigm to see if children could distinguish relative clauses from conjoined clauses, despite the claim of Tavakolian (1981) that they systematically impose a conjoined clause analysis on relatives. In this study, seventeen 3- and 4-year-olds responded to relative clause constructions like (24).

(24) The cat is holding hands with a man who is holding hands with a woman.

(25) The cat is holding hands with a man and is holding hands with a woman.

This sentence was associated with a pair of pictures, one that was appropriate to it and one that was appropriate to the superficially similar conjoined sentence (25). Seventy percent of the 3-year-olds' responses and 94 of the 4-year-olds' responses matched sentences with the appropriate picture rather than with the one depicting the conjoined clause interpretation.

A second technique we tried used a 'silliness' judgment task (see Hsu, 1981) to establish whether children can differentiate relative clauses from conjoined clauses. Ninety-one percent of the responses of the twelve 3- and 4-year-olds tested categorized as 'silly' sentences such as (26), although sentences such as (27) were accepted as sensible 87% of the time.

(26) The horse ate the hay that jumped over the fence.

(27) The man watched the horse that jumped over the fence.

Notice that sentence (26) would not be anomalous if the *that* -clause were misinterpreted as an *and* -clause, or if it were interpreted as extraposed from the subject NP; in both cases, the horse would be the understood subject of the relative clause. The results therefore indicate that most children interpret the *that*-clause in this sentence correctly, i.e., as a subordinate clause modifying *the hay*. Informal testing of adults suggests that the only respect in which children and adults differ on the interpretation of relative clauses is that the adults are somewhat more likely to accept the extraposed relative analysis as well, though even for adults this analysis is much less preferred.

A third experiment, on the phrase structure of relative clause constructions, indicates that children, like adults, treat a noun phrase and its modifying relative clause as a single constituent, inasmuch as they can construe it as the antecedent for a pronoun such as *one*.

In a picture verification study, fifteen 3- to 5-year-olds responded to the instructions in (28).

(28) The mother frog is looking at an airplane that has a woman in it. The baby frog is looking at one too. Point to it.

Ninety-three percent of the time the subjects chose the picture in which the baby frog was looking at an airplane with a woman in it, in preference to the picture in which the baby frog was looking at an airplane without a woman in it. That is, the relative clause was included in the noun phrase assigned as antecedent to the pronoun.

In short: the weight of evidence now indicates that children grasp the structure and meaning of relative clause constructions quite early in the course of language acquisition, as would be expected in view of the central position of these constructions in natural language.

2.2.2. Temporal Terms. Another line of research has yielded support for the claim that presupposition failure is implicated in children's poor linguistic performance. These studies employed sentences containing temporal clauses with *before* and *after*, as in (29).

(29) Push the red car to me before/after you push the blue car.

Clark (1971) and Amidon and Carey (1972) have claimed that most normal, 3- to 5-year-olds do not understand these sentences appropriately. Since Amidon and Carey established that the children were familiar with concepts of temporal sequence (e.g., as expressed by words like *first* and *last*), the implication is that the structure of these adverbial clauses is beyond the scope of the child's grammar at this age.

However, the acting-out tasks employed in these studies were once again unnatural ones which ignored the presuppositional content of the test sentences. Felicitous usage of sentence (29) demands that the pushing of the blue car has already been contextually established by the hearer as an intended, or at least probable, future event; but this was not established in these experimental tasks. It is very likely, then, that these studies underestimated children's ability to comprehend temporal subordinate clauses. For example, Amidon and Carey reported that five and six year old children who were not given any feedback frequently failed to act out the action described in the subordinate clause. Johnson (1975) found that four and five year old children correctly acted out commands such as those in (30) only 51% of the time; again, the predominant error was failure to act out the action described in the subordinate clause.

(30) a. Push the car before you push the truck. (S1 before S2)

b. After you push the motorcycle, push the bus. (After S1, S2)

c. Before you push the airplane, push the car. (Before S2, S1)

d. Push the truck after you push the helicopter. (S2 after S1)

Crain (1982) satisfied the presupposition of the subordinate clause by having the subordinate clause act correspond to an intended action by the subject, and observed a striking increase in children's performance. To satisfy the presupposition, children were asked, before each command, to choose a toy to push on the next trial. The child's intention to push a particular toy was incorporated into the command that was given on that trial. For instance, sentence (30d) could be used felicitously for a child who had expressed his intent to push the helicopter. Correct responses (i.e., responses in which both the main clause and subordinate clause action were performed, and in the correct order) were produced 82% of the time. Crain's interpretation of these results was that the children's improved performance was due to the satisfaction of the presupposition of the subordinate clause.

However, we now note that the results of that study are open to another interpretation. It may be that improved performance was not due specifically to the contextual appropriateness of the sentence, but to the fact that the child's task was simplified because he was provided with more advance information concerning what his task would be. In the act-out or 'do-what-I-say' paradigm applied to temporal terms, the child must discern two aspects of the command: (i) which two toys to move, and (ii) in which order to move them. If the child has established his intent to move a particular toy, his task involving (i) is simplified. Thus, improved performance may be due to the satisfaction of presuppositions or it may be due to the additional information the child possesses.

Another study was conducted to disentangle these two factors (Gorrell, Crain, & Fodor, 1989). In this study, there were four groups of subjects. One group, the Felicity Group (F), was given commands containing *before* and *after* with prior information about the subordinate clause action, just as in the previous experiment. A second group, the Information Group (I) received prior information about the main clause action; note that this does not satisfy the presupposition of the sentence. There was also a third group, the No Context Group (NC), who received no advance information at all, and a fourth group, the Felicity plus Information Group (FI), who received information over and above what would satisfy the felicity conditions since they chose both actions in advance. Consider, for example, a subject in the F group. He would be asked to choose a toy to push. If he chose the bus, for example, a typical command would be (31).

(31) Push the car before you push the bus.

On the other hand, a subject in the I group who had chosen the bus would be given the command (32).

(32) Push the bus before you push the car.

Fifty-six children participated in the study, ranging in age from 3;4 to 5;10 (mean 4;5). Each child was assigned to one of the four groups, which were of equal size and approximately matched for age. The 'game' equipment consisted of 6 toy vehicles arranged in a row on a table between the child and the experimenter. The stimulus set consisted of 12 commands spoken by the experimenter which the child was to act out. There were three sentences of each of the four types illustrated in (30) above. We were careful to balance order of choice with order of action and assignment to clause type.

The results showed a significant difference between the F and FI groups on one hand, and the I and NC groups on the other. Table 1 shows the percentages of correct responses, where a correct response consisted of performing both actions in the sequence specified by the sentence.

Table 1.

		Main Clause Information				mean
		+		-		
Subordinate	+	FI	80%	F	74%	77%
Clause Information	-	I	51%	NC	59%	55%
	mean		65%		66%	

Note that the relevant factor is whether subordinate clause information was provided in advance. An analysis of variance confirmed that the mere **amount** of information provided makes no significant difference. The FI group performed better than the F group by only 6 percentage points, which does not approach statistical significance. And the I group performed just a little worse (non-significantly again) than the NC group.

Although our study was not specifically designed to assess age differences, we performed a post hoc breakdown of correct responses by two age groups: under 4;4, and 4;4 and over. The older group, as one would expect, performed somewhat

better than the younger group. What is perhaps most interesting is that the younger group appear to be even more sensitive than the older group to the proper contextual embedding of utterances.[6]

A breakdown of the types of errors that occurred reveals that the predominant errors are (i) acting out the main clause only, and (ii) reversing the correct order of the actions. As noted above for relative clause constructions, acting out the main clause only is a quite reasonable response given that the context failed to satisfy the subordinate clause presupposition. And in fact most of these errors were found in the I and NC groups.[7] Reversals were the most frequent error type in the study though they constituted only 19% of all responses. These errors may reflect a genuine lack of comprehension of either the temporal terms or the relevant syntactic structure. However no child in either the F or FI groups produced a consistent response pattern which would indicate that this was the case, so it seems more likely that these errors were due primarily just to occasional inattention.

The main conclusion we draw from these results is that children, from a very young age, are indeed sensitive to the proper contextual embedding of language. Their performance is facilitated by satisfying the presuppositions of temporal subordinate clauses, and information which does not satisfy the presuppositions does not result in facilitation.

A secondary conclusion is that children do construct the appropriate syntactic structure for sentences with embedded clauses. If the children in our study had failed to distinguish main from subordinate clauses (e.g., by assigning a 'flat' conjunction-type structure to the experimenter's commands), we would not expect to find the difference between the F and I groups we observed. Nor is it plausible to suggest that the children relied upon a structure-independent formula of 'old information precedes new information.' For example, for the F group, the new information was always in the main clause. If children were assuming that old information would be first, we would have expected relatively poor performance from the F group on sentences in which the main clause preceded the subordinate clause. In fact, no such effect was observed.

In sum: once again, the linguistic knowledge of young children, when freed of interfering influences, appears to be quite advanced. Adults have the ability to set aside contextual factors in an unnatural experimental situation, but children, with their more limited cognitive and social skills, apparently do not have this ability. Consequently, they are highly sensitive to pragmatic infelicities. And therefore their linguistic knowledge can be accurately appraised only by tests which include controls to insure that they are not penalized by their knowledge of pragmatic principles.

2.3. Plans

Another possible source of poor performance by children is in formulating the action plans which are needed in order to obey an imperative, or act out the content of a declarative sentence which they have successfully processed and understood. As we use the term, a plan is a mental representation used to guide action. A plan may be simple in structure, consisting of just a list of actions to be performed in sequence; or it may be internally complex, with loops and branches and other such structures now familiar in computer programs.

Formulating a plan is a skill that makes demands on memory and computational resources. In certain experimental tasks, these demands may outweigh those of the purely linguistic processing aspects of the task. So when children perform poorly, it is important to consider the possibility that formulating, storing or executing the relevant action plan is the source of the problem, rather than imperfect knowledge of the linguistic rules or an inability to apply them in parsing the sentence at hand.

2.3.1. Prenominal Modifiers. The first study on plans that we conducted was in response to the claim by Matthei (1982) and Roeper (1972) that 4- to 6-year-olds have difficulty in interpreting phrases such as (33) containing both an ordinal and a descriptive adjective.

(33) the second striped ball

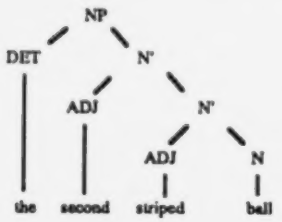
Confronted with an array such as (34), many children selected item (ii), i.e., the ball which is second in the array and also is striped, rather than item (iv) which is the second of the striped balls (counting from the left as the children were trained to do).

(34) Array for "the second striped ball"

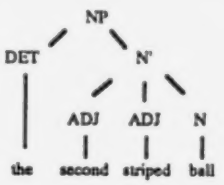
(i) (ii) (iii) (iv) (v)

The empirical finding, then, appears to be that children assign an interpretation that is not the same as an adult would assign to expressions of this kind. This difference is attributed by Matthei to children's failure to adopt the hierarchical phrase structure internal to a noun phrase that characterizes the adult grammar. This structure is shown in (35). Instead, Matthei argues that children adopt a 'flat structure' for phrases of this kind, with both the ordinal and the descriptive adjective modifying the noun directly as in (36).

(35)



(36)



Any divergence between children's and adults' grammars poses a problem from the standpoint of language acquisition theory; namely, explaining how the child ultimately converges on the adult grammar without correction or other 'negative' feedback. Fortunately, there is no need to assume an error in the children's grammar in this case, for there is an alternative component of the language processor in which the errors might have arisen. In a series of experiments, Hamburger and Crain (1984) show that most children do assign the adult phrase structure and do understand the phrase correctly as referring to the second of the striped balls. The difficulty that children experience arises when they attempt to derive from this interpretation a procedure for actually identifying the relevant item in the array. An analysis of the logical structure of the necessary procedure shows it to be quite complex, significantly more so than the procedure for "count the striped balls," the kind of phrase Matthei used in a pretest in an attempt to show that children were able to cope with the nonsyntactic demands of the task.

This procedural account of the children's errors is supported by the sharp improvement in performance that results from three changes in method. One change is the inclusion of a pretask session in which the children handle and count homogeneous subsets of the items which are subsequently used in an array. This experience is assumed to prime some of the procedural planning required in the main experimental task. A second change in method is to withhold the display while the sentence is being uttered, so that formation and execution of the plan are less likely to interfere with each other. A dramatic improvement in performance on a phrase like (33) also results from first asking the child to identify the *first* striped ball, which forces him to plan and execute part of the procedure he will later need for (33). Facilitating the procedural aspects of the task thus makes it possible for the child to reveal his mastery of the syntax and semantics of such expressions.

Hamburger and Crain also found quite direct evidence that children do not assign the 'flat structure' analysis. The standard assumption in linguistics is that proforms corefer with a syntactic constituent. In the correct structure (35), the words *striped* and *ball* form a complete constituent, but in the incorrect structure (36) they do not. Thus the children should permit the proform *one* to corefer with *striped ball* only if they have the correct hierarchical structure. To find out whether they permit this coreference, they were tested on the instructions in (37), with the array in (38).

(37) Point to the first striped ball; point to the second *one*.

(38)

(i) (ii) (iii) (iv) (v)

Hamburger and Crain found that the children consistently responded to the second instruction by pointing to (v) rather than (iv), showing that they took the proform *one* to corefer with expressions like *striped ball*. Thus it appears that they do know the structure (35).

Finally, we note two experiments by Hamburger and Crain (1987). The purpose of these experiments was to provide empirical support for our claim that response planning is an important factor in psycholinguistic tasks, independently of syntax and semantics.

The first experiment attempts to show that children's ability to comprehend a phrase is inversely related to the complexity of the associated plan. For this purpose we compare phrases that are arguably equal to each other in syntactic complexity but differ in plan complexity. Examples are shown in (39), in increasing order of complexity of plans.

(39) i. John's biggest book
ii. second green book
iii. second biggest book

Planning complexity can also be deconfounded from the complexity of semantic constituent processing. Note that semantic considerations lead to the prediction that (39i), would be hardest, because its word meanings have to combine in a way contrary to the surface sequence of words (= the biggest of John's books).

The pattern of children's responses supports the predictions of our procedural account, and does not conform to the account based on syntactic or semantic constituency. Children responded correctly to phrases like (39i) 88% of the time. They gave correct answers only 39% of the time for examples like (39ii), and they were only 17% correct for phrase like (39iii). Thus, this experiment provides clear evidence that plans, not linguistic structures (syntactic or semantic) can determine processing success and failure for young children.

The second experiment addresses the cognitive difficulty of planning by prefacing the test with a sequence of exercises designed to alleviate the planning difficulty. This activity does not provide any extra exposure to the phrases tested; nevertheless, we anticipate a reduction in errors on these phrases. Consider a phrase such as "the second tallest building." This plan requires the interpreter to identify its referent. The child must integrate sequential pairwise comparisons of relative size. In the pre-test activity the child would be shown a display of several objects of one type (say boxes), but of different sizes, and asked to hand the experimenter the biggest one. Then, once this object was removed from the array, the experimenter asked the child to perform the task again, saying, "Now, find the biggest box in this group." In this way the child would identify the second biggest box without ever hearing the phrase "the second biggest box" uttered. Children's comprehension of the phrase was tested before and after the preparatory task. They gave significantly more correct responses (46%) following the preparatory task than before it (8% in this experiment). This result suggests that their difficulty with phrases of this sort stems from the complexity of the response plans.

3. SENTENCE PRODUCTION

To acquire a language is to learn a mapping between potential utterances and associated potential meanings. Successful mastery should reveal itself in both comprehension and production. In the previous section we were concerned with studies of children's comprehension, in which their knowledge is tested by presenting utterances and observing the interpretations that they assign. We now turn to tests of children's competence which proceed in the other direction: the input to the child is a situation, which has been designed to suggest a unique sentence meaning, and the behavior we observe is the utterance by which the child describes that situation.

It would have been reasonable to expect that the sorts of nonsyntactic problems that present obstacles for children in comprehension tasks might prove to be as hard or even harder for them to overcome in production tasks. But we have not found this to be the case. The results of recent elicited production studies are dramatically better than those of comprehension studies directed to the same linguistic constructions. For example, Richards (1976) elicited appropriate uses of the deictic verbs *come* and *go* from children age 4;0 - 7;7, while Clark and Garnica (1974) reported that even 8-year-olds didn't consistently distinguish between *come* and *go* in a comprehension task.

The disparity between production and comprehension studies is particularly striking because it is the reverse of what one would expect. To find production superior to comprehension in children's language is as surprising as it would be to find production superior to comprehension in adult second-language learning, or to find recall superior to recognition in any psychological domain. It is plausible to argue, therefore, that the superiority of production is only apparent, and is due to differences in the sensitivities of production tests and comprehension tests. And the logic of the situation suggests that it is the comprehension tests that are deficient. After all, success is hard to argue with. With suitable controls, successful production by children is a strong indicator of underlying linguistic competence, as long as their productions are as appropriate and closely attuned to the context as adult utterances are. Because there are so many ways to combine words incorrectly, consistently correct combinations in the appropriate contexts are not likely to come about by accident. On the other hand, failure on any kind of psychological task cannot be secure evidence of lack of the relevant knowledge, since the knowledge may be present but imperfectly exploited.

As we saw in the previous section, comprehension studies seem to be particularly susceptible to problems of parsing, planning and so forth which impede the full exploitation of linguistic knowledge. Production tasks appear to be less hampered by these extra-grammatical

factors. This is probably because production avoids non-verbal response planning, which we have seen is a major source of difficulty in act-out comprehension tasks. It is worth noting also that in constructing contexts to elicit particular utterance types, we have no choice but to attend to the satisfaction of the presuppositions that are associated with the syntactic structures in question, because otherwise the subjects won't utter anything like the construction that is being targeted. In elicited production it is delicate manipulations of the communicative situation that give one control over the subject's utterances.

3.1. Relative Clauses. In section 2 we presented evidence of young children's competence with relative clauses. Further confirmation was obtained by Hamburger and Crain (1982), using an elicited production methodology. Pragmatic contexts were constructed in which the presuppositions of restrictive relatives were satisfied. It was discovered that children as young as three reliably produce relative clauses in these contexts.

A context that is uniquely felicitous for a relative clause is one which requires the speaker to identify to an observer which of two objects to perform some action on. In our experiment, the observer is blindfolded during identification of a toy, so the child cannot identify it to the observer merely by pointing to it or saying *this/that one.* Also, the differentiating property of the relevant toy is not one that can be encoded merely with a noun (e.g., *the guard*) or a prenominal adjective (e.g., *the big guard*) or a prepositional phrase (e.g., *the guard with the gun*), but involves a more complex state or action (e.g., *the guard that is shooting Darth Vader*). Young children reliably produce meaningful utterances with relative clauses when these felicity conditions are met. For example:

(40) Jabba, please come over to point to the one that's asleep. (3;5)

Point to the one that's standing up. (3;9)
Point to the guy who's going to get killed. (3;9)

Point to the kangaroo that's eating the strawberry ice cream. (3;11)

Note that the possibility of imitation is excluded because the experimenter takes care not to use any relative clause constructions in the elicitation situation. This technique has now been extended to younger children (as young as 2;8), and to the elicitation of a wider array of relative clause constructions, including relatives with object gaps (e.g., *the guard that Princess Leia is standing on*).

3.2. Passives. Borer and Wexler (1987) have argued that A-chains, which are involved in the derivation of verbal passive constructions, are not available to children in the first few years.[8] Borer and Wexler maintain that knowledge of A-chains is innate, but becomes accessible only after the language faculty undergoes maturational change. We were not convinced, however, that this maturation hypothesis is necessitated by the facts. Rather, the facts seem to be consistent with A-chains being innate and accessible from the outset.

One source of data cited in support of the maturation hypothesis is the absence of full passives in the spontaneous speech of young children. But this of course is not incontrovertible evidence that children's grammars are incapable of generating passives. Full passives are rarely observed in adults' spontaneous speech either, or in adult speech to children. But their paucity is not interpreted in this case as revealing a lack of grammatical knowledge. Instead, it is understood as due to the fact that the passive is a marked form which it is appropriate to use only in certain discourse contexts; in most contexts the active is acceptable and more natural, or a reduced passive without a *by*-phrase is sufficient. That is, the absence of full verbal passives in adult speech is assumed to be a consequence of the fact that it's only in rare situations that the full passive is uniquely felicitous. But the same logic that explains why adults produce so few full passives may apply equally to children. Perhaps they too have knowledge of this construction, but do not use it except where the communicative situation is appropriate.

We have tested this possibility in an experiment with thirty-two 3- and 4-year-old children. (Crain, Thornton, & Murasugi, 1987). One experimenter asked the child to pose questions to another experimenter. The pragmatic context was carefully controlled so that questions containing a full verbal passive would be fully appropriate. The following protocol illustrates the elicitation technique:

Adult: See, the Incredible Hulk is hitting one of the soldiers Look over here. Darth Vader goes over and hits a soldier. So Darth Vader is also hitting one of the soldiers. Ask Keiko which one.

Child to Keiko: Which soldier is getting hit by Darth Vader?

Note that the child knows what the correct answer is to his question, and that he cannot expect to elicit this answer from his interlocutor (Keiko) unless he includes the *by*-phrase. In fact, exactly 50% of responses were passives with full *by* -phrases). Of course, active constructions are also felicitous in this context (e.g., *Which soldier is Darth Vader hitting?*), even though the contextual contrast with another agent (the Incredible Hulk) may tend to favor the passive stylistically. And indeed 31% of responses were active questions with object gaps. The other 19% of responses included mostly sentences that were grammatical but not as specific as the context demanded (e.g., passive lacking *by*-phrases).

Using this technique, we were able to elicit full verbal passives from all but three of the thirty-two children tested so far, including ones as young as 3;4. Some examples are shown in (41).

(41) She got knocked down by the Smurfie. (3;4)
Which girl is pushing, getting pushed by a car? (3;8)
He got picked up from her. (3;11)
It's getting ate up from Luke Skywalker. (4;0)
Which giraffe gets huggen by Grover? (4;9)

Note that these utterances contain a variety of morphological and other errors, but they all nevertheless exhibit the essential passive structure (underlying subject in pre-verbal position; agent in post-verbal prepositional phrase).[9] It might be argued that the children's passives elicited in this experiment do not involve true A-chains. However, since they are just like adult passives (disregarding morphological errors), the burden of proof falls on anyone who holds that adult passives involve A-chains and children's passives do not. No criterion has been proposed, as far as we know, which distinguishes adult's and children's passives in this respect. For example, it is true that the children almost always use a form of *get* in place of the passive auxiliary *be*, but *get* is acceptable in adult passives also. (*Get* is more regular and phonologically more prominent than forms of *be*, and this may be why it is more salient for children.)

Children's considerable success in producing passive sentences appropriate to the circumstances (i.e., their correct pairing of sentence forms and meanings) constitutes compelling evidence of their grammatical competence with this construction. Comparison of these results with the results of testing the same children with two comprehension paradigms (act-out and picture-verification) confirms that, like spontaneous production data, these measures underestimate children's linguistic knowledge.

The finding that young children evince mastery of the passive obviates the need to appeal to maturation to account for its absence in early child language. Maturation cannot of course be absolutely excluded; but a maturation account is motivated only where a construction is acquired surprisingly late—where this means later than would be expected on the basis of processing complexity, pragmatic usefulness in children's discourse, and so forth. (Also, as noted in section 1, some important cross-language and cross-construction correlations need to be established to confirm a maturational approach; see Borer & Wexler, 1987, on comparison of English passives with passive and causative constructions in Hebrew.) The elicited production results suggest that the age at which passive is acquired in English falls well within a time span that is compatible with these other factors, and so maturation does not need to be invoked.

3.3. *Wanna* contraction. Another phenomenon that can be shown by elicitation to appear quite early in acquisition is *wanna* contraction in English. The facts are shown in (42) and (43).

(42) a. Who do you want to help?
b. Who do you *wanna* help?

(43) a. Who do you want to help you?
b. Who do you *wanna* help you?

Every adult is (implicitly) aware that contraction is admissible in (42b) but not (43b). However, on the usual assumption that children do not have access to 'negative data' (i.e., are not informed of which sentences are ungrammatical) it is difficult to see how this knowledge about the ungrammaticality of sentences like (43b) could be acquired from experience (at any age). So this is yet another candidate for innate linguistic knowledge. (What is known innately would be that a trace between two words prevents them from contracting together. The relevant difference between (42b) and (43b) is that in (43b) the *who* is the subject of the subordinate clause and has been moved from a position between the *want* and the *to*. The trace of this noun phrase that is left behind blocks the contraction. In (42b), by contrast, the trace is in object position after *help*, and therefore is not in the way of the contraction.)

Crain and Thornton (in press) used the elicited production technique to encourage children to ask questions that would reveal violations like (43b) if

these were compatible with their grammars. The target productions were evoked by having children pose questions to a rat who was too timid to talk to grown-ups. The details of the procedure are illustrated in the following scenarios:

Protocol for Object Extraction

Experimenter: The rat looks hungry. I bet he wants to eat something. Ask him what.

Child: What do you wanna eat?

Protocol for Subject Extraction

Experimenter: One of these guys gets to take a walk, one gets to take a nap, and one gets to eat a cookie. So one gets to eat a cookie, right? Ask Ratty who he wants.

Child: Who do you want to eat the cookie?

Using this technique, questions involving both subject and object extraction were elicited from 21 children, who ranged in age from 2;10 to 5;5, with an average age of 4;3. The preliminary findings of the experiment are clearly in accord with the expectations of the innateness hypothesis, although we must verify our own subjective assessment of these data using a panel of judges.[10] In producing object extraction questions (which permit contraction in the adult grammar), children gave contracted forms 59% of the time and uncontracted forms 18% of the time. (There were 23% of other responses not of the target form, such as *What can you eat to see in the dark?*) By contrast, children's production of subject extraction questions (where contraction is illicit) contained contracted forms only 4 of the time and uncontracted forms 67% of the time (with 29% of other responses).

The systematic control of this subtle contrast could perhaps have been shown on the basis of spontaneous production data, but the crucial situations (particularly those that call for subject extraction questions) probably occur quite rarely in children's experience, just as they do in the case of the full passive. So it is not easy to gather data in sufficient quantity for statistical analysis. By contrast, the elicitation technique is obviously an efficient way of generating data, and thus facilitates testing for early acquisition of a variety of constructions relevant to the innateness hypothesis.

4. CONCLUSION

In this paper we have reviewed a great many empirical studies. The thread that ties them together is the idea that, when performance problems are minimized in testing situations, children show early knowledge of a wide range of basic constructions. As early as 1965, Bellugi suggested that children's errors on Wh-questions were due, not to a lack of knowledge of the two relevant transformations (Wh-movement and Subject/Auxiliary Inversion), but to a not yet fully developed capacity to apply both rules in the same sentence derivation. Our work extends this general idea to a broader set of linguistic phenomena. Our particular emphasis has been constructions which linguistic theory predicts should require little or no learning because they involve principles which are universal and hence innate. Our findings suggest that the innateness hypothesis for language is still secure even in its simplest form (in which different innate principles are not timed to mature at different developmental stages). Maturation of nonlinguistic abilities appears to be sufficient to account for the time course of linguistic development.

REFERENCES

Amidon, A., & Carey, P. (1972). Why five-year-olds cannot understand *before* and *after*. *Journal of Verbal Learning and Verbal Behavior, 11*, 417-423.

Bellugi, U. (1965). The development of interrogative structures in children's speech. In K. Riegel (Ed.), *The development of language functions*. University of Michigan Language Development Program, Ann Arbor, Report No. 8.

Borer, H., & Wexler, K. (1987). The maturation of syntax. In T. Roeper & E. Williams (Eds.), *Parameter setting*. Dordrecht, Holland: D. Reidel Publishing Company.

Chomsky, N. (1971). *Problems of knowledge and freedom*. New York: Pantheon Books.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Clark, E. V. (1971). On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior, 10*, 266-275.

Clark, E. V., & Garnica, O. K. (1974). Is he coming or going? On the acquisition of deictic verbs. *Journal of Verbal Learning and Verbal Behavior, 15*, 559-572.

Crain, S. (1982). Temporal terms: Mastery by age five. *Papers and Reports on Child Language Development, 21*, 33-38.

Crain, S., & Fodor, J. D. (1984). On the innateness of Subjacency. *Proceedings of the Eastern States Conference on Linguistics* (Vol. 1). Columbus: Ohio State University.

Crain, S., & McKee, C. (1985) Acquisition of structural restrictions on anaphora. *Proceedings of the North Eastern Linguistic Society, 16*. Amherst: University of Massachusetts.

Crain, S., & Nakayama, M. (1987). Structure-dependence in grammar formation. *Language, 63*, 522-543.

Crain, S., Thornton, R., & Murasugi, K. (1987). Capturing the evasive passive. Paper presented at the 12th Annual Boston University Conference on Language Development.

Crain, S., & Thornton, R. (in press). Recharting the course of language acquisition: Studies in elicited production. In N. Krasnegor, D. Rumbaugh, R. Schiefelbusch, & M. Studdert-Kennedy (Eds.), *Biobehavioral foundations of language development*. Hillsdale, NL: Lawrence Erlbaum Associates.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 6*, 291-325.

Goodluck, H. (1986). Children's interpretation of pronouns and null NPs: Structure and strategy. In P. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development* (2nd ed.). Cambridge, MA: Cambridge University Press.

Goodluck, H., & Tavakolian, S. (1982). Competence and processing in children's grammar of relative clauses. *Cognition, 8*, 389-416.

Gorrell, P., Crain, S., & Fodor, J. D. (1989). Contextual information and temporal terms. *Journal of Child Language, 16*, 623-632.

Hamburger, H., & Crain, S. (1982). Relative acquisition. In S. Kuczaj (Ed.), *Language development* (Volume II, pp. 245-274). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hamburger, H., & Crain, S. (1984). Acquisition of cognitive compiling. *Cognition, 17*, 85-136.

Hamburger, H., & Crain, S. (1987). Plans and semantics in human processing of language. *Cognitive Science, 11*, 101-136.

Hsu, J. R. (1981). *The development of structural principles related to complement subject interpretation.* Doctoral dissertation, The City University of New York.

Jakubowicz, C. (1984). On markedness and binding principles. *Proceedings of the Northeastern Linguistic Society*, Amherst, Massachusetts.

Johnson, H. (1975). The meaning of *before* and *after* for preschool children. *Journal of Experimental Child Psychology, 19*.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition, 2*, 15-47.

Lasnik, H., & Crain, S. (1985). On the acquisition of pronominal reference. *Lingua, 65*, 135-154.

Lust, B. (1981). Constraint on anaphora in child language: A prediction for a universal. In S. Tavakolian (Ed.), *Language acquisition and linguistic theory* (pp. 74-96). Cambridge, MA: MIT Press.

Matthei, E. M. (1981). Children's interpretations of sentences containing reciprocals. In S. Tavakolian (Ed.), *Language acquisition and linguistic theory* (pp. 97-115). Cambridge, MA: MIT Press.

Matthei, E. M. (1982). The acquisition of prenominal modifier sequences. *Cognition, 11*, 301-332.

Nakayama, M. (1987). Performance factors in subject-aux inversion by children. *Journal of Child Language, 14*, 113-125.

Otsu, Y. (1981). *Universal grammar and syntactic development in children: Toward a theory of syntactic development.* Unpublished doctoral dissertation, M.I.T.

Phinney, M. (1981) The acquisition of embedded sentences and the NIC. *Proceedings of the North Eastern Linguistic Society, 11*. Amherst: University of Massachusetts

Richards, M. M. (1976). Come and go reconsidered: Children's use of deictic verbs in contrived situations. *Journal of Verbal Learning and Verbal Behavior, 15*, 655-665.

Roeper, T. W. (1972). *Approaches to a theory of language acquisition with examples from German children.* Unpublished doctoral dissertation, Harvard University.

Roeper, T. W. (1986). How children acquire bound variables. In B. Lust (Ed.), *Studies in the acquisition of anaphora, Volume I*. Dordrecht, Holland: D. Reidel Publishing Company.

Solan, L. (1983). Pronominal reference: Child language and the theory of grammar. Dordrecht, Holland: D. Reidel Publishing Company.

Solan, L., & Roeper, T. W. (1978). Children's use of syntactic structure in interpreting relative clauses. In H. Goodluck & L. Solan (Eds.), *Papers in the Structure and Development of Child Language. UMASS Occasional Papers in Linguistics* (Vol. 4, pp. 105-126).

Tavakolian, S. L. (1978). Children's comprehension of pronominal subjects and missing subjects in complicated sentences. In H. Goodluck & L. Solan (Eds.), *Papers in the Structure and Development of Child Language. UMASS Occasional Papers in Linguistics* (Vol. 4, pp. 145-152).

Tavakolian, S. L. (1981). The conjoined-clause analysis of relative clauses. In S. Tavakolian (Ed.), *Language acquisition and linguistic theory* (pp. 167-187). Cambridge, MA: MIT Press,.

de Villiers, J. G., & de Villiers, P. A. (1986). The acquisition of English. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition Volume I: The data.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wexler, K., & Chien, Y. (1985) The development of lexical anaphors and pronouns. *Papers and Reports on Child Language Development*. Stanford, CA: Stanford University.

FOOTNOTES

**Language and cognition: A developmental perspective* (in press). Norwood, NJ: Ablex.

†Also University of Connecticut, Storrs.

††Also Graduate Center, City University of New York.

[1]Susan Carey has pointed out (Boston University Conference, 1988) that the linguistic maturation hypothesis predicts that knowledge of a linguistic principle should correlate with gestation age rather than with birth age in children born prematurely. Unfortunately, variability is probably such that no clear correlation could be expected to show itself by age 4 or 5, when passives and other relevant syntactic constructions are claimed to emerge.

[2]For full details of procedure and results of this experiment, and of all other studies reported in this paper, we refer readers to the original publications.

[3]As far as is known at present, no natural language exhibits this blanket prohibition against backward pronominalization; see discussion in Lasnik and Crain (1985). This suggests that it is not a possible constraint in a natural language grammar, in which case it should not be entertained by children at any age or stage of acquisition (unless one assumes linguistic maturation).

[4]The awkwardness of the prosodic contour for (18), with its heavy juncture before the final word, may indicate that this kind of construction is also an unnatural one for the sentence production routines.

[5]In reviewing the literature on relative clauses, de Villiers and de Villiers (1986) suggest that if earlier work had counted the assertion-only response as correct, children would have been seen to perform better there too. This objection is unwarranted, for two reasons. First, responses of this type did not appear in other studies, presumably because these studies failed to meet the presuppositions of the restrictive relative clause. More important, in the Hamburger and Crain study this response was not evinced by any of the 3-year-old children, and accounted for only 13% of the responses of the 4-year-olds. Nevertheless, even the 3-year-olds acted out sentences with relative clauses at a much higher rate (69%) of success than in earlier studies.

[6]These results should be interpreted with caution due to the small and unequal number of subjects in each subgroup leading to rather uneven data. For example, the older F group performed relatively poorly compared to the younger F group,

though closer analysis reveals that this is due to the poor performance of just one child (4;4) in the F group.

[7]There were no main-clause-only errors in the FI group. For the F group, 12 of the 16 main-clause-only errors (out of 168 responses) are due to one child.

[8]An A-chain is the association of a trace with a moved noun phrase in an A-position (= argument position such as Subject). For example, in *The bagel was eaten by Bill* there is an A-chain consisting of *the bagel* and its associated trace after *eaten*.

[9]The proper reversal of underlying subject and object order occurred even when the task was complicated by an implausible scene to be described. For example, the sentence *One dinosaur's being eated from the ice cream cone* was used to describe a situation in which the dinosaur was indeed being eaten by the ice cream, not vice versa.

[10]In preliminary evaluation of the audio tapes, we have found it unexpectedly easy to distinguish children's contracted and non-contracted forms in most cases.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 137-147

Cues to the Perception of Taiwanese Tones*

Hwei-Bing Lin† and Bruno H. Repp

A labeling test with synthetic speech stimuli was carried out to determine to what extent the two dimensions of fundamental frequency (Fo), height and movement, and syllable duration provide cues to tonal distinctions in Taiwanese. The data show that the high level vs. mid level tones and the high falling vs. mid falling tones can be reliably distinguished by Fo height alone, whereas the distinction between tones with dissimilar contours, such as the high falling and low rising tones, is predominantly cued by Fo movement. However, the other dimension of Fo may collaborate with the dominant one in cueing a tonal contrast, depending on the extent to which the two tones differ along that dimension. Syllable duration has a small additional effect on the perception of the distinction between falling and nonfalling tones. These results are consistent with previous findings in tone languages other than Taiwanese in that they suggest that tones are mainly cued by Fo. While the primacy of Fo dimensions as cues to tonal contrasts depends on the contrast to be distinguished, the present findings show that tones which nominally differ only in register (e.g., high falling vs. mid falling) may exhibit perceptually relevant contour differences, and vice versa.

INTRODUCTION

The primary acoustic attribute distinguishing linguistic tones is their fundamental frequency (Fo), although duration and amplitude of the syllable carrying the tone may also exhibit characteristic differences. This observation raises the question of whether, in perceiving a phonemic tone, listeners integrate all these acoustic cues, or whether they pay attention to Fo alone.

Several studies have investigated the role of Fo versus other properties as cues to tonal distinctions, in both synthetic and natural speech. For example, Abramson (1962) imposed artificial Fo movements on natural Thai monosyllables by means of a vocoder and found that Fo overpowered other concomitant features such as duration and amplitude in cueing tonal distinctions. The primacy of Fo was confirmed in a later study using synthetic Thai speech (Abramson, 1975), though addition of natural amplitude contours improved identification further. The conclusion that Fo carries sufficient information for conveying tonal distinctions has also been drawn by Howie (1976), Tseng (1981), and M.-C. Lin (1987), who investigated the tones of Mandarin Chinese. However, these studies either did not vary duration and amplitude at all, or they pitted these dimensions against unambiguous Fo contours. In whispered monosyllables, where Fo is altogether absent but duration and amplitude differences may be retained to some extent, tonal distinctions are resolved poorly, though above chance level (e.g., Abramson, 1972; Howie, 1976). It is conceivable that, in addition to increasing the naturalness of utterances (cf. M.-C. Lin, 1987; Rumyantsev, 1987), duration and amplitude have larger effects on tone identification when Fo provides ambiguous information. Also, the relative informativeness of different acoustic cues for tonal identity may vary across languages.

This research, which formed part of the first author's doctoral dissertation, was supported by NICHD Grant HD-01994 to Haskins Laboratories. Special thanks are due to Arthur Abramson, Carol Fowler, and Ignatius Mattingly for their helpful comments on an earlier version of this paper, and to Jackson Gandour and Eva Gårding for serving as reviewers for *Language and Speech*.

The Fo dimension itself may be decomposed into two aspects: height and movement.[1] The relative importance of these two aspects obviously depends on the nature of the tonal distinction to be made: If the distinction is between two tones differing primarily in register (e.g., "high" vs. "low"), Fo height will be important; if two tones differ primarily in contour (e.g., "rising" vs. "falling"), Fo movement will be the dominant cue. However, for tones that, according to linguistic nomenclature, differ only in register or in contour, the other aspect of Fo might play a secondary role in cueing the contrast. Furthermore, for tones that differ in both register and contour (e.g., "high falling" vs. "low rising"), both Fo height and movement may be relevant, though perhaps not equally important. Their relative importance may depend on what other tones there are in the language.

With these issues in mind, we conducted the present study to determine the relative importance of Fo height, Fo movement, and duration as cues to the tonal distinctions of Taiwanese. From traditional classifications by phonologists and from acoustic studies (Chiang, 1967; H.-B. Lin, 1988; Zee, 1978) it is evident that Taiwanese has five long tones, whose typical Fo contours are illustrated in Figure 1.[2] They fall into two classes: Tones 1 ("high level") and 5 ("mid level") are fairly level or static, whereas tones 2 ("high falling"), 3 ("mid falling"), and 4 ("low rising") are contoured or dynamic. Two pairs of tones, 1 vs. 5, and 2 vs. 3, have similar Fo contours but differ in register. We would thus expect Fo height to play a primary role in the distinction of these tone pairs. As for the tonal pairs with different Fo contours, we expected that Fo movement rather than height would be responsible for cueing the differences. However, we wondered whether the "other" aspect of Fo would make a contribution to a distinction as well. For example, would the small difference in Fo movement between the two level tones, 1 and 5, or the larger one between the two falling tones, 2 and 3, play any role in perception at all? And would Fo height be relevant to the distinction, for instance, between tones 1 and 4, even though they differ in contour? The answers to these questions seemed less obvious.

With respect to duration, the five Taiwanese long tones fall into two groups: relatively long (tones 1, 4, and 5) and relatively short (tones 2 and 3). In isolated syllables, the respective average durations were 145 ms and 75 ms (H.-B. Lin, 1988; see Figure 1). Thus, in principle, duration could be a rather strong cue for the distinction between falling and nonfalling tones.

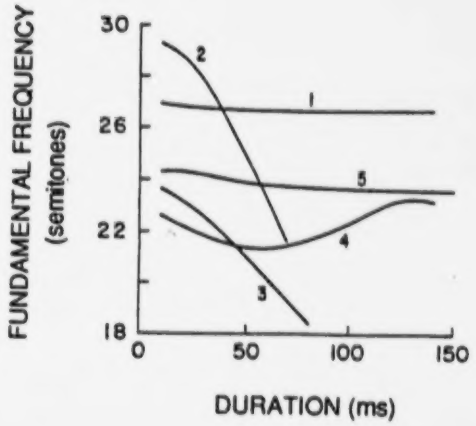


Figure 1. **Average Fo movements of five Taiwanese tones produced by three male speakers on the syllable /do/ in sentence-final position. (Data from H.-B. Lin, 1988.)**

Our approach was to synthesize a variety of Fo patterns by varying Fo height, Fo movement, and duration independently in isolated syllables.[3] In some of our stimuli, the Fo heights, movements, and durations of typical Taiwanese tones were juxtaposed in novel combinations, so the relative strength of these competing cues could be assessed. In addition, we synthesized Fo patterns intermediate between those of the original tones in terms of Fo height and/or movement. These relatively ambiguous stimuli provided the best opportunity to observe effects of secondary cues, such as duration or the "other" Fo dimension, on perception of a given tonal contrast. Even though our stimuli were presented singly, we conceptualized the study in terms of pairwise tonal contrasts, for heuristic convenience. Listeners, of course, always had all five lexical alternatives in mind as they tried to identify our synthetic syllables.

METHODS

Materials

The syllable /do/ was modelled after a natural Taiwanese utterance on the software serial formant synthesizer at Haskins Laboratories. Since the Taiwanese /d/ is unaspirated and voiceless (i.e., [t]), only a 10-ms release burst preceded the onset of voicing. The onset

frequencies of the first three formants were 450 Hz, 1160 Hz, and 2400 Hz. After 70 ms, they reached respective steady states of 560 Hz, 760 Hz, and 2000 Hz. The amplitude of the voicing source was kept at a constant value. The extreme Fo values and durations of the five basic tones are shown in Table 1. The actual Fo movements followed the natural models (cf. Figure 1) as closely as possible.

Table 1. *Average Fo onset and offset values (in Hz) and durations (in ms) of the five synthetic tones on /do/. The "pivot" indicates the turning point of the Fo movement in tone 4.*

Tone	onset	pivot	offset	duration
1	130		129	145
2	150		96	75
3	109		80	75
4	102	94	105	145
5	113		107	145

To assess the relative importance of Fo height, Fo movement, and duration cues in the perception of tonal distinctions, these properties were varied independently within each pairwise contrast. Thus we synthesized, in addition to the original tones, stimuli in which the Fo movement of tone X was combined with the Fo height of tone Y, and vice versa. This involved a translation of the whole Fo movement up or down the linear frequency axis, by adding a positive or negative constant to all Fo values in the synthesis specifications. We defined Fo height operationally as the onset frequency of a tone.[4] In addition, we created an Fo movement intermediate between the two original tonal contours by averaging their Fo values, and we chose an intermediate Fo onset frequency as well. Thus we had three Fo movements (X, Y, and intermediate) and three Fo heights (X, Y, and intermediate), all combinations of which resulted in nine stimuli for any given tonal contrast. The stimulus set for the tone 2-4 contrast is illustrated in Figure 2; tone 24 denotes the stimulus intermediate between tones 2 and 4 in both Fo height and Fo movement.[5]

Given five original tones, there were 10 pairs of tonal contrast. In four of these (2-3, 1-4, 1-5, 4-5), the durations of the two original tones were the same, and so all nine stimuli were synthesized at the same duration. In the six other contrasts (1-2, 1-3, 2-4, 2-5, 3-4, 3-5), the two original tones had different durations (cf. Table 1). For these, we synthesized the set of nine stimuli at three different durations, the two original ones (75 and 145 ms) and one intermediate one (110 ms), which resulted in 27 stimuli. When the duration of an original Fo movement was changed, its onset and offset frequencies were maintained, but the Fo trajectory was stylized by linear interpolation between the extreme values. In the case of tone 4, the location of the pivot was moved in proportion to the duration change, and two linear interpolations were performed.

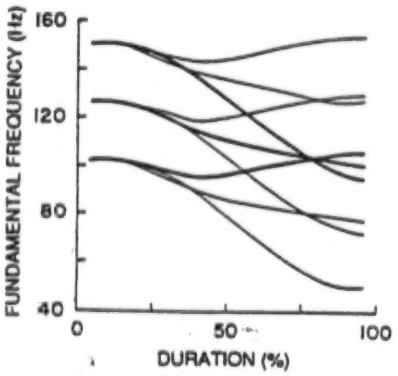


Figure 2. **Example of a stimulus set for a particular tonal contrast: tone 2 versus tone 4. The original (2 and 4) and intermediate (24) Fo movements are shown by the heavy lines; the other patterns were obtained by changing the onset frequencies of these three Fo movements. (See also Footnote 5.)**

In all, 6 X 27 + 4 X 9 = 198 stimuli were created on the synthesizer, though a number of these were identical or closely similar to each other.[6] The stimuli were recorded in five different randomizations, each on a separate audio tape. On each tape, there were six blocks of 33 stimuli, separated by 8 sec of silence. There was a 3.5 sec interstimulus interval within blocks. Before the presentation of the test tapes, subjects had a chance to familiarize themselves with the original synthetic tones on the /do/ syllable. The order of test tapes was counterbalanced across subjects, and the experimental session took about 90 minutes.

Subjects

Four male and four female listeners were recruited from University of Connecticut and Yale University graduate students and were paid for their participation. All were native speakers of Taiwanese and reported to have no history of speech or hearing disorder. Like all educated Taiwanese, they were also fluent in English and Mandarin Chinese.[7]

Procedure

The stimuli were presented binaurally over headphones at a comfortable listening level. Listeners were instructed to identify each /do/ syllable as (1) 'city,' (2) 'gambling,' (3) 'envy,' (4) 'map,' or (5) 'surname' by writing down the number of the choice. The response choices were listed on the answer sheet in Chinese characters. Subjects were told not to leave any blanks.

RESULTS

The results for the 10 tonal contrasts will be discussed in the following order: First, the two contrasts that have nominally identical contours but differ in register (1-5, 2-3); then, two contrasts that have similar Fo movements and differ in Fo height (4-5, 1-4); finally, the remaining six contrasts which have very dissimilar Fo movements (as well as differences in duration), grouped into three pairs exhibiting different magnitudes of Fo height difference (2-1, 2-5; 5-3, 2-3; 3-4, 2-4). In naming the members of a tonal pair, the one with the higher onset frequency is always named first (hence 2-1 and 5-3).

Tables 2 to 11 show the response distributions for the stimuli pertaining to each of the 10 pairs of tonal contrast. In each table, stimuli are coded in terms of Fo movement and height with two-digit numbers standing for the intermediate values on these two dimensions (e.g., 15 is intermediate between the original tones 1 and 5). To simplify the tables, the results have been averaged across stimuli differing in duration; effects of stimulus duration will be discussed later. Thus, each stimulus set includes nine types. The average recognition rate for the five synthetic syllables modelling the original tones was 91% correct, which confirms that these stimuli were satisfactorily synthesized (cf. Footnote 3).

Tones with the same contour, differing in register

Presumably, the more pronounced the difference in an acoustic property between two tones, the more important it will be as a cue to the tonal distinction. If, moreover, differences along other psychoacoustic dimensions are small, it will emerge as the dominant cue. Thus, for pairs such as tones 1 vs. 5 and 2 vs. 3, which nominally have the same Fo contour but differ in register, Fo height was expected to be the dominant cue. However, tones 2 and 3, at least, also exhibit a difference in Fo movement (see Figure 1), and we wondered whether that dimension would contribute to the perceptual distinction.

Tone 1 vs. tone 5. Tones 1 (high level) and 5 (mid level) have almost identical, flat Fo movements; the difference between them is in Fo height. The data in Table 2 reveal that Fo height indeed plays a primary role in the perception of this distinction. Movement 1 with height 5 was identified predominantly as tone 5, whereas movement 5 with height 1 was classified as tone 1. There were virtually no confusions with other tones. Did Fo movement have any cue value at all? The stimuli with intermediate Fo height, which should have been the most sensitive indicators of Fo movement effects, suggest a negative answer. However, movement 1 with height 5 received only 70 tone 5 responses, whereas the original synthetic tone 5 received 90. This difference notwithstanding, the effect of Fo movement on the perception of this tonal distinction seems negligible simply because the difference between the tones is minimal on this dimension.

Table 2. *High level tone (1) vs. mid level tone (5).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
1	1	80				20
	15	57.5			2.5	40
	5	30				70
15	1	80			2.5	17.5
	15	62.5				37.5
	5	2.5		2.5		95
5	1	90			2.5	7.5
	15	52.5		2.5		45
	5	7.5	2.5			90

Tone 2 vs. tone 3. The falling tones 2 and 3 nominally have the same Fo contour, though Fo in tone 2 falls somewhat more steeply than tone 3. The main difference between them is in Fo height. The data in Table 3 confirm that Fo height is the major cue for the distinction: Movement 2 with height 3 was identified mostly as tone 3, and movement 3 with height 2 as tone 2. However, there was also some effect of Fo movement: At each of the three Fo heights, more tone 3 responses were obtained when the movement derived from tone 3 rather than from tone 2. Thus, even though both tones are considered merely "falling" in traditional phonological terminology, there is in fact a perceptually relevant difference in Fo movement between them. The difference in Fo height, however, is clearly the dominant cue.

Table 3. *High falling tone (2) vs. mid falling tone (3).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
2	2		92.5	7.5		
	23		82.5	17.5		
	3		22.5	72.5		5
23	2		82.5	15		2.5
	23		80	17.5		2.5
	3		17.5	80		2.5
3	2	7.5	65	27.5		
	23		67.5	32.5		
	3		7.5	92.5		

Tones with similar contours, differing in register

The contour of tone 4, referred to here as "rising", is actually a complex falling-rising or dipping Fo movement with a relatively limited range (cf. Figure 1). Because of that limited range, it looks somewhat similar to the flat contours of tones 1 and 5, though to the ear it may be quite dissimilar. In contrasting tones 5 vs. 4 and 1 vs. 4, which differ in both Fo register and contour, we expected both dimensions of Fo to be perceptually relevant. The question was which of them would carry more weight.

Tone 4 vs. tone 5. Tones 4 (low rising) and 5 (mid level) are relatively similar in Fo movement during the first half of their durations, but during the second half tone 4 moves up and almost merges with tone 5. There is also a difference in Fo height, tone 4 having a lower onset than tone 5. However, the data in Table 4 reveal Fo movement to be the primary cue:

Table 4. *Mid level tone (5) vs. low rising tone (4).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
5	5	5		5		90
	54	2.5		7.5		90
	4				2.5	97.5
54	5			2.5	65	32.5
	54			2.5	70	27.5
	4				95	5
4	5				100	
	54	2.5			97.5	
	4				100	

High intelligibility of both tones was maintained when their original Fo movements were presented at uncharacteristic heights. An effect of Fo height emerged only when Fo movement was intermediate. Evidently, Fo movement is the dominant cue to this distinction. This is also suggested by the finding that tone 4 responses predominated for the intermediate Fo movement: Detection of even a slight "dip" was sufficient to elicit tone 4 percepts.

Tone 1 vs. tone 4. Since tone 1 (high level) is very similar to tone 5 in Fo movement but higher in register, the Fo movement difference between tones 1 and 4 is similar to that between tones 5 and 4, just discussed, only the difference in height is larger. Does this imply a larger role of Fo height in cueing the distinction? Table 5 suggests an affirmative answer, but it is evident that Fo movement is still the dominant cue:

Table 5. *High level tone (1) vs. low rising tone (4).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
1	1	82.5				17.5
	14	25				75
	4		2.5	2.5	10	85
14	1	75			22.5	2.5
	14	7.5			67.5	25
	4				100	
4	1	15			82.5	2.5
	14			2.5	97.5	
	4				100	

Movement 4 with height 1 was still predominantly identified as tone 4, and movement 1 with height 4 was identified as tone 5, which shares the contour with tone 1. A small effect of Fo height is evident with the original tone 4 movement: At a very high Fo, some tone 1 responses did occur. A large effect of Fo height was obtained for the intermediate movement stimuli, whose identification changed from tone 1 to tone 4 as the height was lowered. This shift is larger than that observed in Table 4, in accord with the larger height difference for the present contrast.

Tones with dissimilar contours

The falling Fo contour (tones 2 and 3) is acoustically and perceptually very dissimilar from the level and rising contours; it is also carried on a shorter syllable, though differences in duration will be ignored for the time being. Since Fo

movement proved to be perceptually important even for distinguishing tones with relatively similar contours, we certainly expected that tonal contrasts involving falling tones would be primarily cued by Fo movement, with secondary contributions of Fo height depending on the amount of the difference. In fact, it will be seen that, because for each falling tone and each level tone in Taiwanese there is a similar tone differing in register, stimuli with altered Fo height were often identified as tones other than those in the particular contrast under consideration.

Tone 2 vs. tone 1. Tones 2 (high falling) and 1 (high level), though they are both nominally "high", do show a difference in onset frequency. The most striking difference, however, is in their Fo movements. As the data in Table 6 show, Fo movement is indeed the dominant cue to the contrast: Movement 2 with height 1 was still mostly identified as tone 2, whereas movement 1 with height 2 was recognized as tone 1. The stimuli with the intermediate Fo movement showed some effects of Fo height on their identification as tone 1, but the effect was such that tone 1 responses increased as the height was raised, even though the higher onset frequency derived from tone 2; there was no effect of height on tone 2 responses. The paradoxical effect of Fo height derived from a tendency to identify the intermediate Fo movement as tone 3, which indeed has a contour intermediate between tones 2 and 1, and occasionally even as tone 5; both of these tendencies to give mid-register tone responses increased as Fo was lowered, at the expense of tone 1 responses. Basically, Fo height seems to be irrelevant to the perception of the tone 2 vs. 1 contrast.

Table 6. *High falling tone (2) vs. high level tone (1).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
2	2		96.6	3.3		
	21		95	5		
	1		87.5	12.5		
21	2	25.8	55	18.3		0.8
	21	12.5	61.6	22.5		3.3
	1	6.6	59.1	25.8		8.3
1	2	96.6		1.6		1.6
	21	98.3				1.6
	1	92.5		0.8		6.6

Tone 2 vs. tone 5. The difference in Fo movement between tones 2 (high falling) and 5 (mid level) is the same as that between tones 2 and 1, just discussed, but the difference in Fo height (onset) is much larger. The data presented in Table 7 show that predictable confusions occurred as Fo height was changed: Movement 2 with height 5 was often identified as tone 3, while movement 5 with height 2 was obviously tone 1. Stimuli with intermediate movement were predominantly identified as falling tones (tones 2 or 3, depending on Fo height), though at a very high Fo (above the characteristic height of tone 1) some tone 1 responses occurred. In general, however, Fo movement remained the overriding cue for this falling versus level tone distinction.

Table 7. *High falling tone (2) vs. mid level tone (5).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
2	2		96.6	3.3		
	25		83.3	14.1	0.8	1.6
	5		34.1	60		5.8
25	2	27.5	51.6	17.5	1.6	1.6
	25	4.1	61.6	28.3	0.8	5
	5		32.5	60.8	1.6	5
5	2	97.5	1.6	0.8		
	25	86.6		2.5		10.8
	5	6.6	0.8	5		87.5

Tone 5 vs. tone 3. Here we have another level versus falling contrast, at a lower Fo height. Tones 3 (mid falling) and 5 (mid level) originate at almost the same frequency, so no effect of Fo height was expected. The data in Table 8 confirm that the dominant cue for the tone 5 versus tone 3 distinction is Fo movement, though the stimuli with the intermediate movement do show a small effect of Fo height. The copies of the original tones were not identified very well in this pair of tones; this reflects in part their inherent confusability with tones having the same contour (tones 1 and 2, respectively) but, in addition, the neutralization of duration cues and stylization of Fo movements may have increased the number of confusions. This general ambiguity may have made listeners extra sensitive to small differences in Fo height.

Table 8. *Mid level tone (5) vs. mid falling tone (3).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
5	5	11.6	1.6	33	0.8	82.5
	53	4.1		58	1.6	88.3
	3	1.6		7.5	2.5	88.3
53	5	1.6	10	49.1	1.6	37.5
	53		6.6	65.8		27.5
	3		7.5	65.8		26.6
3	5		23.3	70.8		5.8
	53		15.8	80	0.8	3.3
	3		20	67.5		12.5

Tone 1 vs. tone 3. Tones 1 (high level) and 3 (mid falling) not only differ in Fo movement, as do tones 5 and 3, but also in Fo height. As can be seen in Table 9, subjects' responses changed considerably with both changes in Fo movement and in Fo height. However, changes in height led to predictable "confusions": Movement 1 (level) with height 3 (mid) was largely identified as tone 5 (mid level), whereas movement 3 (falling) with height 1 (high) was most often classified as tone 2 (high falling). These responses thus occurred merely reflect the fact that Fo height cues the tone 1 vs. tone 5 and tone 2 vs. tone 3 distinctions. The intermediate movement stimuli, however, do reveal a genuine effect of Fo height on the contrast between tones 1 and 3: Tone 1 responses decreased and tone 3 responses increased as height was lowered, with an increase in tone 5 confusions in the middle. Thus it appears that both Fo height and movement are important cues for the distinction between tones 1 and 3, just as we expected.

Table 9. *High level tone (1) vs. mid falling tone (3).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
1	1	85		1.6	1.6	11.6
	13	63.3		3.3		33.3
	3	5	0.8	6.6	9.1	78.3
13	1	31.6	23.3	12.5	0.8	31.6
	13	6.6	12.5	28.3		52.5
	3		7.5	56.6	0.8	35
3	1	3.3	69.1	24.1		3.3
	13		45	50		5
	3		15	81.6		3.3

Tone 3 vs. tone 4. There is a striking difference in Fo movement between tone 3 (mid falling) and tone 4 (low rising). However, the difference in Fo onset is small. The data in Table 10 confirm that Fo movement is the primary cue for the distinction between these tones: Small changes in Fo height left the responses to the original Fo movements unchanged. The stimuli with the intermediate movement did show an effect of Fo height, despite the relatively small physical differences involved. However, the effect was less on identification of these stimuli as tones 3 or 4, but primarily on their identification as tone 5 (mid level). Indeed, the intermediate movement was relatively flat and thus could be mistaken for the mid level tone when its height was raised. For the tone 3 vs. tone 4 distinction, therefore, Fo height seems to be of little importance.

Table 10. *Mid falling tone (3) vs. low rising tone (4).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
3	3		14.1	80.8		5
	34		9.1	85		5.8
	4		10	83.3		6.6
34	3	1.7	1.7	23.3	5	68.3
	34	1.7	0.8	38.3	23.3	35.8
	4	0.8		37.5	33.3	28.3
4	3			3.3	96.6	
	34			5	95	
	4			2.5	97.5	

Tone 2 vs. tone 4. Tones 2 (high falling) and 4 (low rising) show the sharpest Fo movement contrast of any tone pair, as well as the largest difference in Fo height. As can be seen in Table 11, movement 2 with height 4 was identified as tone 3 (mid falling), which is not surprising. Interestingly, however, movement 4 with height 2 was identified as tone 1 (high level), even though movement 4 with height 1 was not so identified (see Table 5). Thus, the movement barrier between tones 1 and 4 can be overcome by a sufficient raising of Fo height. The intermediate movement stimuli (somewhat falling with a dip) were never labeled as tone 4, but were highly ambiguous at a high Fo and perceived as tone 3 at a low Fo. Clearly, both Fo movement and height are important for the tone 2 versus tone 4 distinction, though the actual weights of these cues are difficult to gauge because of the intrusion of other responses.

Table 11. *High falling tone (2) vs. low rising tone (4).*

Stimuli Mov	Hgt	Responses (%) 1	2	3	4	5
2	2	0.8	95	4.1		
	24		72.5	25		2.5
	4		11.6	82.5		5.8
24	2	37.5	28.3	20.8		13.3
	24	0.8	32.5	44.1		22.5
	4		3.3	86.6		10
4	2	86.6	1.6	0.8	10.8	
	24	7.5		4.1	88.3	
	4	0.8		4.1	95	

Overview of Fo effects

In the preceding diacussion of pairwise tonal contrasts, stimuli included in one particular subset were often also relevant to the perception of other tones, as evidenced by the various response intrusions. It is useful, therefore, to survey the pattern of predominant identification responses for the complete set of stimuli, still disregarding variations in duration. Table 12 provides such an overview. In its columns, Fo heights are arranged in terms of decreasing onset frequencies, and in its rows Fo movements are ordered in terms of decreasing differences between the onset and offset frequencies (with one minor, deliberate reversal). Each original height (movement) occurred with 9 different movements (heights), whereas each intermediate height (movement) occurred with 3 different movements (heights). For each height-movement combination, Table 12 lists all tonal categories with more than 20 of responses, in rank order.

The first five columns show that strongly falling Fo movements were perceived as either tone 2 or tone 3, depending on Fo height. The secondary relevance of Fo movement to this distinction can be seen in the fact that tone 3 responses increased as the steepness of the Fo movement decreased. The next four columns show that moderately falling Fo movements were perceived as tone 3 or tone 5 at the lower Fo onsets; stimuli with high onsets were not well sampled here, but suggest tone 1 responses with tone 2 as the second choice. The next three columns show that shallow falling Fo movements (with an onset-offset difference of 6 Hz or less) were invariably identified as level tones: as tone 1 or as tone 5, depending on Fo height. Finally, the last three columns show that, when the curvature of tone 4 is imposed on a flat Fo movement, the stimuli were mostly heard as tone 4. This salient Fo movement cue was overridden only by very high absolute Fo values, which favored tone 1 percepts.

Table 12. *Predominant response categories (with percentages exceeding 20%) for all combinations of Fo movements and heights. Numbers in parentheses represent onset frequencies (for Fo height) and the difference between onset and offset frequencies (for Fo movement).*

Fo height Type	Onset (Hz)	Fo movement Type and onset-offset difference (Hz) 2 (45)	23 (41.5)	25 (30)	3 (29)	12 (27.5)	24 (25.5)	35 (17.5)	13 (15)	34 (13)	5 (6)	15 (3.5)	1 (1)	45 (1.5)	14 (-1)	4 (-3)
2	(150)	2	2	2,1	2,3	2,1	1,2,3	—	—	—	1	—	1	—	—	1
12	(140)	2	—	—	—	2,3	—	—	—	—	—	—	1	—	—	—
25	(131.5)	2	—	2,3	—	—	—	—	—	—	1	—	—	—	—	—
1	(130)	2	—	—	2,3	2,3	—	—	1,5,2	—	1	1	1	—	1,4	4
23	(129.5)	2	2	—	2,3	—	—	—	—	—	—	—	—	—	—	—
24	(126)	2,3	—	—	—	—	3,2,5	—	—	—	—	—	—	—	—	4
15	(121.5)	—	—	—	—	—	—	—	—	—	1,5	1,5	1,5	—	—	—
13	(119.5)	—	—	—	3,2	—	—	—	5,3	—	—	—	1,5	—	—	—
14	(116)	—	—	—	—	—	—	—	—	—	—	—	5,1	—	4,5	4
5	(113)	3,2	—	3,2	3,2	—	—	3,5	—	—	5	5	5,1	4,5	—	4
35	(111)	—	—	—	3	—	—	3,5	—	—	5	—	—	—	—	—
3	(109)	3,2	3	—	3	—	—	3,5	3,5	5,3	5	—	5	—	—	4
45	(107.5)	—	—	—	—	—	—	—	—	—	5	—	—	4,5	—	4
34	(105.5)	—	—	—	3	—	—	—	—	3,5	—	—	—	—	—	4
4	(102)	3	—	—	3	—	3	—	—	3,4,5	5	—	5	4	4	4

Duration as a cue

Finally, we consider the possible role of duration as a cue to tonal distinctions. Because the two falling tones, 2 and 3, are characterized by shorter durations than the other tones (see Table 1), a short stimulus duration was expected to be a cue to the falling contour category. In addition, shortening the duration of a falling Fo movement increases its slope, which may further enhance the "falling" percept. This may favor tone 2 over tone 3 responses for clearly falling contours, considering the steeper slope of the tone 2 movement (cf. Figure 1).

The results were analyzed by comparing the response percentages for the short (75 ms) and long (145 ms) durations of each stimulus whose duration was varied. (The intermediate durations were not considered.) Table 13, which is arranged in the same way as Table 12, lists all response categories in which a change of more than 10 occurred as stimulus duration was shortened. (The largest change was 35.) Plus signs indicate increases, minus signs decreases, in order of absolute magnitude. The changes are often complementary for two tonal categories, with one response increasing at the expense of another.

Our expectations were that falling tone (2 and 3) responses would generally increase and level tone (1 and 5) responses would decrease as a consequence of stimulus shortening, but that for strongly falling Fo movements tone 2 responses might increase at the expense of tone 3 responses. The overall pattern of changes supports the first prediction: There were 11 increases in tone 2 responses versus 3 decreases, 13 increases in tone 3 responses versus 3 decreases, one increase in tone 5 responses versus 12 decreases, and no increase in tone 1 responses versus 3 decreases. The second, more specific prediction was less well supported: In the left half of Table 13, increases in tone 2 responses are frequent, but there are also some decreases, and changes in tone 3 responses are inconsistent, though usually complementary to the changes in tone 2 responses. On the whole, it appears that duration did have a role as a secondary cue for the falling-nonfalling tone distinction.

DISCUSSION

From these results, it is quite clear that Fo is the most prominent perceptual cue for tonal contrasts in Taiwanese, as in all other tone languages studied so far. The present data further show that either dimension of Fo (height or movement) can emerge as the dominant factor in cueing a particular tonal contrast, depending on the tonal patterns to be differentiated.

Table 13. *Response categories showing changes in excess of 10 following a change in stimulus duration from 145 to 75 ms. All relevant combinations of Fo heights and movements are shown.*

Fo height		Fo movement										
Type	Onset (Hz)	Type and onset-offset difference (Hz)										
		2 (45)	25 (30)	3 (29)	12 (27.5)	24 (25.5)	35 (17.5)	13 (15)	34 (13)	5 (6)	1 (1)	4 (-3)
2	(150)	—	—	—	+2	-1,+2,+3,-5	—	—	—	—	—	—
12	(140)	-2,+3	—	—	+2,-3	—	—	—	—	—	—	—
25	(131.5)	—	—	—	—	—	—	—	—	-1	—	—
1	(130)	+2	+2	+2,-3	+2	—	—	-5,+2,+3	—	—	—	—
23	(129.5)	—	—	—	—	—	—	—	—	—	—	—
24	(126)	—	—	—	—	+2,-5	—	—	—	—	—	-4,+3
15	(121.5)	—	—	—	—	—	—	—	—	—	—	—
13	(119.5)	—	—	—	—	—	—	-3	—	—	-1,+5	—
14	(116)	—	—	—	—	—	—	—	—	—	—	—
5	(113)	+3,-2	+2	+2	—	—	-5,+3	—	—	-5	—	—
35	(111)	—	—	-2,+3	—	—	-5,+3	—	—	-5	—	—
3	(109)	—	—	—	—	—	-5,+3	-5,+3	-5+3	—	-5,+4	-4,+3
45	(107.5)	—	—	—	—	—	—	—	—	—	—	—
34	(105.5)	—	—	—	—	—	—	—	-5	—	—	—
4	(102)	—	—	—	—	—	—	—	—	—	—	+3

In general, the more pronounced an acoustic difference between two tones is, the more likely it will be an important cue in perceiving the contrast. Thus, Fo height was found to be a main cue in the distinction between tones with highly similar contours but different registers, whereas distinctions between tones with dissimilar contours were mainly cued by Fo movement.

On the whole, Fo movement seems to be perceptually more important than Fo height for Taiwanese listeners. This is especially evident in the contrasts between tones 1 and 4, and tones 5 and 4, which in principle could have depended on Fo height. One reason for this finding may be that Fo movement is a more stable dimension than Fo height, which varies across speakers and is often ambiguous when no contextual reference is provided. (Note that tones 2 and 3, and tones 1 and 5, tend to be confused in isolated syllables.) However, it has also been observed that linguistic background can affect the perception of Fo patterns (Gandour, 1978, 1983). Thus it could be that Fo height and movement are given different weights in languages with different tonal inventories, even when similar tonal contrasts are being perceived. For instance, in Cantonese, Fo height is rather important because four out of six tones are relatively similar in Fo movement (Vance, 1977). On the other hand, all four tones of Mandarin are dissimilar in Fo movement (Howie, 1976). Taiwanese represents an intermediate case. Gandour (1983) found that Cantonese speakers rely more heavily on Fo height, and less on Fo movement, than do speakers of Mandarin and Taiwanese when judging various Fo patterns. Our finding of the relative importance of Fo movement in the perception of Taiwanese tones is not inconsistent with Gandour's findings.

One purpose of perceptual studies such as the present one is to go beyond linguistic nomenclature, which omits acoustic detail, and beyond phonetic studies, which describe but do not establish the perceptual relevance of these detailed aspects. Thus we have shown that the Taiwanese high falling and mid falling tones, which nominally differ only in register, also exhibit some perceptually relevant differences in Fo movement. We have also shown that the striking difference in duration between falling and nonfalling tones can provide distinctive information in ambiguous cases. Bailey and Summerfield (1980), in their thorough studies of the perception of segmental phonetic distinctions, have argued that any systematic difference in acoustic properties between segments can be shown to be perceptually relevant. This generalization also seems to apply to the perception of tonal distinctions.

REFERENCES

Abramson, A. S. (1962). *The vowels and tones of standard Thai: acoustical measurements and experiments.* Bloomington: Indiana University Research Center in Anthropology, Folklore, and Linguistics.

Abramson, A. S. (1972). Tonal experiments with whispered Thai. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 31-44). The Hague: Mouton.

Abramson, A. S. (1975). The tones of Central Thai: Some perceptual experiments. In J. G. Harris & J. R. Chamberlain (Eds.), *Studies in Tai Linguistics in honor of William J. Gedney* (pp. 1-16). Bangkok: Central Institute of English Language.

Bailey, P. J. & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performances, 6,* 536-563.

Chiang, H. T. (1967). Amoy-Chinese tones. *Phonetica, 17,* 100-115.

Gandour, J. (1978). The perception of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 41-76). New York: Academic Press.

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics, 11,* 149-175.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones.* Cambridge, UK: Cambridge University Press.

Lin, H.-B. (1988). *Contextual stability of Taiwanese tones.* Doctoral dissertation, University of Connecticut, Storrs.

Lin, M.-C. (1987). The perceptual cues of tones in standard Chinese. *Proceedings of the Eleventh International Congress of Phonetic Sciences, Vol. 1* (pp. 162-165). Tallinn, Estonia, U.S.S.R.: Academy of Sciences of the Estonian SSR.

Rumyantsev, M. K. (1987). Chinese tones and their duration. *Proceedings of the Eleventh International Congress of Phonetic Sciences, Vol. 1.* (pp. 166-169). Tallinn, Estonia, U.S.S.R.: Academy of Sciences of the Estonian SSR.

Tseng, C. Y. (1981). *An acoustic phonetic study of tones in Mandarin Chinese.* Doctoral dissertation, Brown University, Providence, RI.

Vance, T. J. (1977). Tonal distinctions in Cantonese. *Phonetica, 34,* 93-107.

Zee, E. (1978). Duration and intensity as correlates of Fo. *Journal of Phonetics, 6,* 221-225.

FOOTNOTES

Language and Speech, 32, 25-44 (1989).

†Also Department of Linguistics, University of Connecticut, Storrs. Now at the Lexington Center, Inc., Jackson Heights, New York.

[1]Throughout this paper, we refer to Fo characteristics of phonological tones as Fo *register* and *contour*, but to the corresponding phonetic dimensions as Fo *height* and *movement*. Register and contour thus are characterized in discrete terms (high, mid, low; rising, level, falling), whereas height and movement are continuously variable and are described in acoustic terms. The term "register" is not intended to denote changes in vocal register.

[2]According to traditional classification, Taiwanese has five long tones and two short tones. The former occur in syllables ending with vowels or nasals, whereas the latter occur in checked syllables only (see Chiang, 1967). Because only an open syllable was used as the carrier in the present study, short tones were excluded from consideration.

[3]We did not vary amplitude characteristics of the stimuli, to keep the design within bounds. To confirm that Taiwanese tones could be identified in isolated syllables, tones produced on the

syllable /do/ by a native speaker were presented to native listeners for identification in a pilot study. Although confusions occurred between tones 2 and 3, and between tones 1 and 5, average performance was 87 correct.

[4]As a result, variations in Fo height were restricted for the tone 3-5 and 4-5 contrasts. Alternatively, we could have chosen the Fo midpoint or the average Fo as our measure of Fo height. In that case, however, height variations would have been restricted for the tone 1-2, 2-3, and 4-5 contrasts. Our choice of Fo onset frequency as the relevant measure is consistent with phonological terminology for tones, which usually mentions onset register and direction of contour, such as "high falling".

[5]Figure 2 actually shows a tonal pair of different original durations, whose Fo movements have been scaled to a common duration. The figure is slightly inaccurate because it shows curvilinear Fo movements for both original tones. In reality, after two temporally contrasting tones had been scaled to a common duration, one or both of the Fo movements were linear.

[6]For example, each original tone occurred four times because it occurred in four different stimulus subsets (i.e., in contrast with four other tones). The slightly different response percentages to physically identical stimuli in different tables (below) derive from the fact that each stimulus subset was treated separately in the data analysis.

[7]Note that /do/ is not a possible syllable in Mandarin, so second-language interference seemed unlikely.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 148-161

Physical Interaction and Association by Contiguity in Memory for the Words and Melodies of Songs*

Robert G. Crowder,† Mary Louise Serafine,† and Bruno H. Repp

Six experiments investigated two explanations for the *integration effect* in memory for songs (Serafine et al., 1984, 1986). The integration effect is the finding that recognition of the melody (or text) of a song is better in the presence of the text (or melody) with which it had been heard originally than in the presence of a different text (or melody). One explanation for this finding is the *physical interaction* hypothesis, which holds that one component of a song exerts subtle but memorable physical changes on the other component, making the latter different from what it would be with a different companion. Experiments 1, 2, and 3 investigated the influence that words could exert on the subtle musical character of a melody. A second explanation for the integration effect is the *association-by-contiguity* hypothesis—that any two events experienced in close temporal proximity may become connected in memory such that each acts as a recall cue for the other. Experiments 4, 5, and 6 investigated the degree to which both successive and simultaneous presentations of spoken text with hummed melody would give rise to association of the two components. The results gave encouragement for both explanations and are discussed in terms of the distinction between encoding specificity and independent associative bonding.

Stimuli obviously have multiple features. Two examples are that ordinary objects have both color and shape and that songs have both melody and text. Questions about memory representations of these theoretically separable but seemingly related components of a song—melody and text—motivated our earlier investigations (Serafine, Crowder, & Repp, 1984; Serafine, Davidson, Crowder, & Repp, 1986). We hypothesized that a song might be represented in memory in three ways: (1) *independent storage* of components (the separate entities perceived and stored so that memory for one is uninfluenced by the other); (2) *holistic storage* (the two components so thoroughly connected in perception and memory that one is remembered only in the presence of the other); and (3) *integrated storage* (the two components related in memory such that one component is better recognized in the presence of the other than otherwise). The holistic hypothesis is obviously false in the general case since people often recognize the melodies of familiar songs when they hear them performed on solo instruments, or with unfamiliar verses. What this informal observation leaves open, however, is whether the memory representation consists of independent or integrated components.

In earlier studies we reported evidence for what we called an *integration effect* in memory for melody and text. Using a recognition task, we found that melodies were better recognized when heard with the same words (as originally heard) than with different words, even when the different words fit the melody and were equally familiar to the subject. Similarly, we found that the words of songs were better recognized in test songs containing the original melody than in those containing a different but equally familiar melody. The procedure we employed was as follows: Subjects heard a serial presentation of up to 24

This research was supported by NSF Grant GB 86 08344 to R. Crowder and by NICHD Grant HD01994 to Haskins Laboratories. The authors are grateful for the assistance of William Flack in testing subjects and to Shari R. Speer for discussion of earlier versions of this paper.

unfamiliar folksong excerpts, each heard only once. A recognition test followed immediately, in which subjects were typically asked to indicate, for each excerpt, whether they had heard exactly that melody (or text) before, ignoring the current text (or melody). The test excerpts consisted of old songs (exactly as heard in the presentation) and various types of new songs (for example, old melody with new words), including a type we termed mismatch songs—that is, an old melody with old words that had been paired with a *different* melody in the original presentation. The critical comparison, of course, was between melody recognition when old songs were heard and when mismatch songs were heard, that is, when the melody was paired with its original companion as opposed to a different, but equally familiar, companion. This comparison, then, avoided the potentially biasing effect that completely new, unfamiliar words would have on recognition of a truly remembered melody. What we have termed the integration effect is the finding that both melody and text recognition were better in the case of old songs than in mismatch songs. We concentrated on the facilitating effect of identical words on recognition of melodies because recognition of words was in some cases almost at ceiling.

The effect was robust. It was not eliminated by instructing the subjects, on their initial hearing, to focus attention on the melody (and ignore the words), nor was it eliminated by hearing a different singer on the recognition test than had sung in the original presentation (Serafine et al., 1984). Moreover, the effect was not due to a particular experimental artifact, the potentially confusing effect of hearing the melody with seemingly "wrong" words; the wrong words did not make melody recognition suffer, as against an appropriate baseline, but the right words facilitated it (Serafine et al., 1986).

The integration was not accounted for by a semantic connotation imposed on the melody by the meanings of the words (Serafine et al., 1986) because the integration effect was found even in songs employing nonsense syllables on presentation and test. A melody heard only once, then, was better recognized in the presence of its original nonsense text than with different but equally familiar nonsense. This latter observation seems inconsistent with a *meaning-interaction* hypothesis that might have considerable intuitive appeal.

In the present studies, we explore further the source of the integration effect. Two hypotheses, not necessarily incompatible, are under test here, the *physical interaction hypothesis* and the *association-by-contiguity*[1] *hypothesis*. The first of these asserts that when a song is sung, the words impose subtle effects upon the melody notes, slightly affecting their acoustic properties such as the onsets, durations, and offsets. We have termed these effects "submelodic" because they would leave unaffected the pitches as they would be notated or conceived in composition. For example, some words might impose a *staccato* articulation and others a *legato* phrasing. If this hypothesis were correct, then a melody sung with one particular text would in fact be a somewhat different melody than it were when sung with another text. It would not be surprising to find that the melody were then better recognized with the same words both times than with changed words. A similar argument could be made for texts.

The association-by-contiguity hypothesis asserts that two events that occur in close temporal proximity (contiguously or simultaneously) tend to be associated in memory, though neither was necessarily changed by virtue of having entered into this association. If this hypothesis were correct, then in the limit text and melody would be just as well associated if they were experienced simultaneously but separately (e.g., words spoken and hummed melodies) as if they were given as a song.

In the present research, the first three experiments addressed the *submelodic* hypothesis (a special case of the physical-interaction hypothesis in which words affect musical properties of the melody) and the latter three experiments addressed the association hypothesis. All experiments employed our usual general procedure: Subjects heard folksong excerpts followed immediately by a melody recognition test where test items contained controlled combinations of song components. All experiments used variations of the musical materials and design described below.

General Method

Musical materials were based on 40 American folksongs (from Erdei, 1974, see Serafine et al., 1984, 1986) which, in earlier experiments, we found were virtually all unfamiliar to our subjects. There were 20 pairs of song excerpts, each pair selected so that melodies and texts were interchangeable, having rhythmic compatibility. Figure 1 shows such a pair.

Interchangeability of melodies and texts was crucial to the construction of test items in which a song contained a different melody or text than that heard originally in the presentation. Thus, each text contained a stress pattern suitable for either melody, and both texts within a pair contained the same number of syllables. The exceptions were Song Pairs 11 and 17, where one text was shorter by one syllable and required the common "slur" across two tones (see "sleep" in Figure 1, Melody B). Given interchangeable components, each pair potentially yielded five types of test items, examples of which are shown in Figure 2:

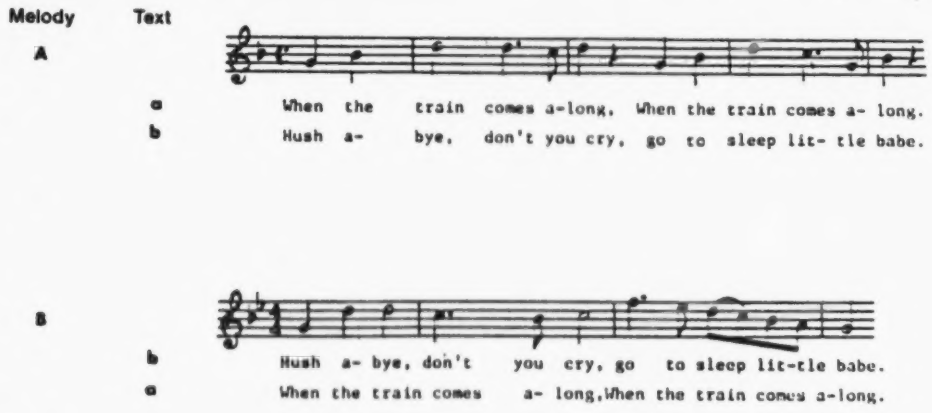

Melody
Text
A
a
When the train comes a-long, When the train comes a- long.
b
Hush a- bye, don't you cry, go to sleep lit- tle babe.
B
b
Hush a- bye, don't you cry, go to sleep lit-tle babe.
a
When the train comes a- long,When the train comes a-long.

Figure 1. **Sample pairs of songs with interchangeable texts. (Aa and Bb denote original songs; Ab and Ba denote derivatives.)**

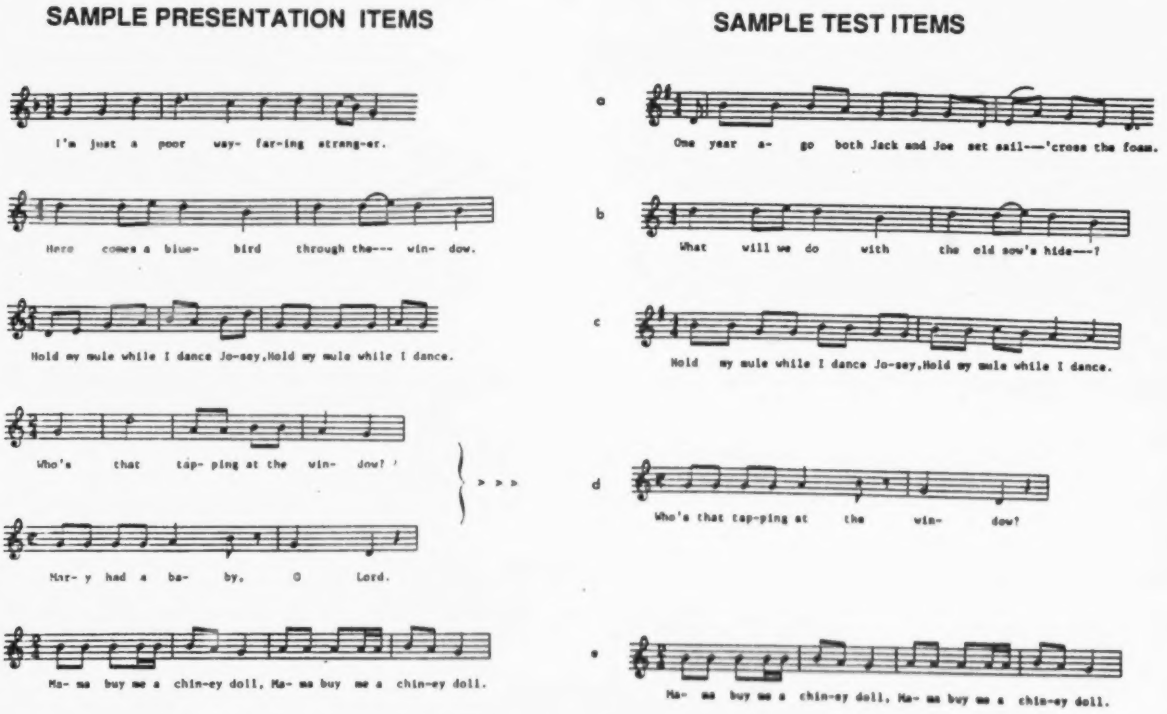

SAMPLE PRESENTATION ITEMS
SAMPLE TEST ITEMS
I'm just a poor way- far-ing strang-er.
Here comes a blue- bird through the--- win- dow.
Hold my mule while I dance Jo-sey,Hold my mule while I dance.
Who's that tap- ping at the win- dow?
Mar- y had a ba- by, O Lord.
Ma- ma buy me a chin-ey doll, Ma- ma buy me a chin-ey doll.
a
One year a- go both Jack and Joe set sail---'cross the foam.
b
What will we do with the old sow's hide---?
c
Hold my mule while I dance Jo-sey,Hold my mule while I dance.
d
Who's that tap-ping at the win- dow?
e
Ma- ma buy me a chin-ey doll, Ma- ma buy me a chin-ey doll.

Figure 2. **Sample presentation and test items (Code for test items: a - new melody/new words, b - old melody/new words, c- new melody/old words, d- old melody/old words, mismatched, e- old songs).**

The five types were a - new melody/new words, b - old melody/new words, c- new melody/old words, d- old melody/old words, mismatched, e- old songs.

In addition, test items might consist of old or new melodies alone, that is, hummed melodies without words. The present experiments employed three to five types of test items, but in each case the critical comparison was between old songs and mismatch songs, where the latter items allowed us to test recognition of one component in the presence of a different component that had nevertheless been heard in the presentation and was equally familiar.

The songs were sung in the alto range by the second author, recorded onto a master tape and dubbed onto sets of experimental tapes with a 5-sec interval of silence between presentation items and a 10-sec response interval after each test item. A silent metronome set at one beat per sec facilitated performance at an even tempo, and a piano tone (not heard by the subjects) ensured pitch accuracy at the start of each song.

Subjective tempos across the songs were not uniform, however, due to normal rhythmic and metric variations (e.g., "double time"). The presented songs ranged in total duration from about 4 to about 10 sec, with a mean of 6.4 and standard deviation of 2.01. All songs used C as the tonic, although there were variations in mode (Dorian, major and minor), and starting tone. Only slight alterations were made in the original folk melodies or texts (e.g., "across" changed to "cross"), in order to ensure rhythmic interchangeability of materials.

The same general design was used in all experiments. The presentation and test sequences always utilized the song pairs in the same order. On the presentation tapes, half the songs were melodies with their original folksong texts (type Aa in Figure 1) and half used the borrowed, interchangeable text (type Ab in Figure 1). Each mismatch item on the test tapes required two songs in the presentation sequence (since the melody of one would be tested with the text of the other). Whenever two such songs occurred in the presentation, they followed one another immediately on the tape. Natural sources of variation among these songs include length, nature of the melody, tempo, and subject matter of the text, to name only a few characteristics. These factors were completely controlled, however, by counterbalancing across different subjects groups.

Experiment 1

The aim of Experiment 1 was to test the submelodic hypothesis by employing, on the test tape, songs that contained different texts but the same phonetic and prosodic pattern as those employed on the original presentation. We derived phonetically similar texts by translating each text into a corresponding nonsense text, where vowels were left intact and consonants were changed to a reasonably close phonetic neighbor (/b/=/g/; /k/=/t/, etc.). The presentation consisted of songs with nonsense texts, and the test consisted of obviously different but phonetically similar texts, that is, the real words. If the submelodic hypothesis were correct, the integration effect should be obtained. That is, a melody should be better recognized when it appears with words that are phonetically similar to the nonsense words with which it was originally heard than with words whose phonetic derivatives are equally familiar but had been heard originally with a different melody.

Subjects heard a presentation of 24 songs with nonsense texts followed by a 20-item melody recognition test containing four each of the following types of items:

(a) "old songs" (old melody with real words that are phonetically similar to the nonsense text heard with that melody in the presentation).[2]
(b) "mismatch songs" (old melody with real words that are phonetically similar to a nonsense text heard with a different melody in the presentation);
(c) new melody/"old words" (new melody with real words that are phonetically similar to a nonsense text heard in the presentation);
(d) old melody hummed, (that is, without words);
(e) new melody hummed.

The main question was whether melody recognition would be better in the "old song" than in the "mismatch" condition. The hummed test items provided a baseline for melody recognition.

Method

Materials

Using the songs described under General Method, each text was translated into phonetically similar nonsense using the rules described in Experiment 1 of Serafine et al. (1986). The following are examples of translated texts written following English orthography:

Original: Cobbler, cobbler, make my shoe.
Nonsense: Togglue, Togglue, nate nie choo.
Original: Cape Cod girls they have no combs.
Nonsense: Tade top berf shey jaze mo tong.

Design

Five sets of presentation and test tapes were constructed, each administered to a different group of subjects. Presentations consisted of 24 song excerpts with nonsense texts, and test sequences consisted of 20 items containing real words, where each of the five types of items occurred 4 times. Across five subject groups, each presentation item was tested in each of the five conditions. For example, the first presentation item was tested as an "old song" in one group, as a "mismatch song" in another group, as an old melody hummed in another group, and so on. Because each "mismatch song" required hearing two songs in the presentation, the presentation tapes consisted of 20 song excerpts plus 4 additional ones for the "mismatch" condition.

Procedure

Testing was conducted individually in a quiet laboratory with tapes heard over loudspeakers. Subjects were instructed to listen to a presentation of 24 excerpts that would sound like folksongs except that the texts had been changed to nonsense. They were told that their "memory would be tested later" but not informed that only melody recognition would be tested. The test immediately followed the presentation. Subjects were told that test items would consist of hummed melodies or songs with real words, but in all cases they were only to indicate whether they had "heard that exact melody before--that is, just the musical portion." Subjects indicated "yes" or "no" on the answer sheet and gave a confidence rating that ranged from 1 to 3.

Subjects

Twenty Yale undergraduates with undetermined levels of musical training were equally divided among the five groups.

Results and Discussion

Yes/no responses with confidence ratings were translated into single scores with a theoretical range of 1 to 6, where 1 represents *very confident no* (did not hear melody) and 6 represents *very confident yes* (did hear melody). The mean ratings for "old song," "mismatch song," new melody "old words," old hummed melody, and new hummed melody, respectively, were 3.99, 4.44, 3.56, 4.00, and 2.92. An analysis of variance with subjects as the sampling variable showed conditions to be a significant source of variance, $F(4,76)=9.45$, $p < .001$. The Newman-Keuls procedure was used to identify which comparisons produced the significant overall effects. Evidence that melody recognition was above chance was provided by the fact that the mean rating for old melodies hummed (4.00) exceeds that for new hummed melodies (2.92), $p < .01$, as well as by the fact the rating of "mismatch songs" (4.44) was reliably greater than the condition with new melodies and "old words" (3.56). However, "mismatch songs" generated a higher mean rating (4.44) than did "old songs" (3.99), contrary to our expectations based on the submelodic interpretation of the integration effect. Thus, phonetically similar "old song" texts did not enhance recognition for the original melodies.

In this experiment, then, the submelodic hypothesis was not supported. No evidence emerged that the nonsense words in the presentation imposed effects on the melodies which would allow the phonetically similar real words to enhance melody recognition. As a test of the submelodic hypothesis, however, this experiment seemed in retrospect to have been compromised by the fact that subjects heard real words on the test after having heard nonsense in the presentation. Possibly the surprise of a full semantic experience after originally studying nonsense may have been distracting. In Experiments 2 and 3 we addressed this problem.

Experiment 2

One aim of the present experiment was to verify that the integration effect could be obtained with newly-constructed nonsense texts that would be necessary for Experiment 3, where the submelodic hypothesis was tested again. In an earlier study (Experiment 1 of Serafine et al., 1986) we had shown, as noted above, that the integration effect was robust with nonsense texts of the sort used in Experiment 1 of the present paper. In the present and following experiments, however, somewhat different rules were employed for the construction of nonsense texts, and we sought to verify that the resulting new materials would give rise to the integration effect. Following a presentation of song excerpts with nonsense words, a melody recognition test employed only three types of test items:

(a) old songs (old melody, old nonsense words) exactly as heard in the presentation);
(b) mismatch songs (old melody with old nonsense words that had been sung to a different melody in the presentation);
(c) new melody/old words.

The critical comparison was that between the old song and mismatch conditions. The main prediction was that melody recognition would be enhanced by the presence of the original old nonwords (in the old songs) over that obtained with the different but equally familiar nonwords (in the mismatch songs).

Method

Materials

The design of Experiment 2 required only eighteen of the 20 song pairs we had available. (Two pairs were omitted on the grounds that they had proven at least somewhat familiar to some subjects in Experiment 1). Because, in Experiment 3 reported below, *two* nonsense texts were required for each real text, it was necessary to employ new rules for translating real words into nonsense. The new rules, as follows, were similar to, but more detailed than, those of Experiment 1 and allowed fewer deviations. For example, the voiced/unvoiced distinction was was preserved across transformations:

(1) Vowels remain the same, and the following vowel-liquid sequences are treated as intact : /ɛr/ as in *Mary*, /ar/ as in *far*, /ɪl/ as in *will*, /ɔr/ as in *lore*, /ɔɹ/ as in *boy*, /aʊ/ as in *how*, /eɪl/ as in *pail*, /al/ as in *doll*, /ɔl/as in *awl*, /ʌnz/ as in *runs*.

(2) Consonants are interchanged according to the following list of phonetic similarities. For example, if /b/ occurs in a real word, the two corresponding nonsense words use /d/ and /g/, respectively:

/b/=/d/=/g/
/n/=/m/=/l/
/r/=/l/=/j/ or /w/; /w/=/j/=/r/ or /l/
/p/=/t/=/k/
/f/=/θ/=/h/ or /s/; /θ/ =/f/=/ʃ/ or /s/; /h/=/θ/ =/f/; /s/= /ʃ/=/f/
/z/=/v/=/ð/
/tr/=/kw/=/pl/
/pr/=/tw/=/kl/
/kr/=/tw/=/pl/
/st/=/sk/=/sp/
/sl/=/ʃw/=/fr/
/skw/=/str/=/spl/
/bl/=/dw/=/gr/
/sp/=/st/=/sk/
terminal /n/=terminal /m/=terminal /ŋ/

Special cases of translated vowel/consonant combination:

/ɚl/=/ɚm/=/ɚn/ (e.g., *girl = berm, dern*)
/ir/=/il/=/in/ (e.g., *here = seal=feen*)
/end/=/ɛmd/ =/ɛmb/
/ʌn/=/ʌm/=/ʌ l/
terminal /ɔŋ/=/ɔn/=/ɔm/ (e.g., *song= fawn, shawm*)
terminal /ɪŋ/ = /ɪm/ = /ɪn/

(3) Interior /ɚ/ or /ɚn/ is treated as a vowel, but in terminal position is interchanged as follows: /ɚ/=/ɔn/=/ɔl/ as in *anger=often=able*.

(4) A terminal /s/ or /z/, when a plural marker, may be retained (untranslated) if the resulting nonsense is too difficult to pronounce.

(5) The sounds /tʃ/ and /dʒ / are omitted from all real texts because three suitable phonetic correspondences do not exist. Thus, minor changes in some real texts were made (e.g., *Joe* changed to *Moe*, *chase* to *run*).

The following is an example of a translated text, written in the form (regular orthography) used by the singer:

Original: Cobbler, cobbler make my shoe
Nonsense 1: Poggrel, poggrel nate nie foo.
Nonsense 2: Toddwen, toddwen lape lie thoo.

Only one set of nonsense texts was used in this experiment.

Design

The design was comparable to that of Experiment 1. Three sets of presentation and test tapes were administered to different sets of subjects. Across the three subject groups, each presentation item was tested in each of the three conditions: old song, mismatch, and new melody/old words. The presentation tapes consisted of 24 songs, and test sequences consisted of 18 items, six each of the three conditions.

Procedure

The testing procedure was comparable to that of Experiment 1. Subjects were instructed to listen to a presentation of folksong excerpts with nonsense texts, were told that their "memory would be tested later," and following the presentation were given the melody recognition test in which they were to indicate whether they had "heard this exact melody before--that is, just the musical portion." They were not told what types of items to expect on the test except that the nonsense folksongs would be similar to those on the presentation.

Subjects

Fifteen Yale undergraduates with undetermined levels of musical training were equally divided among the three groups.

Results and Discussion

As in Experiment 1, responses were translated into 6-point ratings where 1 represents *very confident no* (did not hear melody) and 6 represents *very confident yes* (did hear melody). Mean ratings for the old song, mismatch, and new melody/old words conditions were 4.85, 3.63, and 2.59 respectively. The results of two omnibus analyses of variance were significant: With subjects as the sampling variable, $F(2,28)=42.66$, $p < .001$, and with the 18 test items as the sampling variable, $F(2,34)=41.76$, $p < .001$. Newman-Keuls tests revealed that melody recognition was significantly better in the old song than in the mismatch condition, both across subjects ($p < .01$) and across items ($p < .01$).

Thus the integration effect was confirmed with these new materials, verifying that the presence of original old words—even nonsense words absent of semantic meaning—facilitates melody recognition over that obtained with the different but equally familiar words, in the mismatch songs. Besides vindicating our new stimuli, the results of Experiment 2 provide welcome replication for one of our most important previous results: In this new experiment, mean ratings for the old song and mismatch conditions, respectively, were 4.85 and 3.63; corresponding means from Experiment 1 of Serafine et al. (1986) were 4.47 and 3.76.

Experiment 3

The aim of Experiment 3 was to retest the submelodic hypothesis, which had not been confirmed in Experiment 1. To avoid the potentially distracting use of both nonsense (at presentation) and real words (at test), which may have influenced the outcome of Experiment 1, we employed *two different sets* of phonetically derived nonsense texts, based on the same real words (which were never used in this experiment). The presentation consisted of folksong excerpts with nonsense texts. The test consisted of folksong excerpts whose texts were phonetically similar to those in the presentation but nevertheless were, in all cases, different nonsense. (As in Experiment 1, the phonetic derivative of an old song was called an "old song," etc.) Test items were of three types:

(a) "old songs" (old melody with nonsense words phonetically similar to the old nonsense text);

(b) "mismatch songs" (old melody with nonsense words phonetically similar to an old nonsense text from a different song in the presentation);

(c) new melody/"old words" (new melody with nonsense words phonetically similar to an old nonsense text).

If the submelodic hypothesis were correct, that is, if words impose subtle and memorable effects upon their melodies, then a melody should be better recognized when it is heard with nonsense words that are phonetically related to the nonsense with which that melody was originally presented than when heard with nonsense that is not phonetically related to the original. In other words, melody recognition in "old songs" should exceed that in "mismatch songs."

Method

Materials

The materials were those described under Experiment 2. Both sets of nonsense texts, (phonetic derivatives of the original folksong texts) were employed.

Design

The design was comparable to that of Experiment 2, except that the test items, instead of comprising old songs, mismatch songs, and new melodies with old words, used the phonetically derived "old songs," "mismatch songs," and new melodies with "old words," where our quotation marks indicate that *exact* repetition of the verbal texts between learning and test never occurred. As in earlier experiments, counterbalancing across subjects groups was employed to control for natural variations in the songs. The presentation consisted of 24 items and the tests consisted of 18 items.

After 12 of the 30 subjects had been tested, an inadvertent error in the test tapes was detected. Two song pairs contained faulty material for the condition new melody/"old words," although the other two conditions were correct. Thus, scores for those 12 subjects were based on four (instead of six) items in the new melody "old words" condition.

Procedure

The procedure was analogous to that of Experiments 1 and 2. At test, subjects were told that the texts of songs may sound similar to or different from those heard before, but they were to attend only to the melody and indicate recognition (yes or no) and a confidence rating on the answer sheet.

Subjects

Thirty adults with undetermined levels of musical training were paid to participate and were equally divided among the three groups.

Results and Discussion

As before, melody recognition ratings had a possible range of from 1 to 6. Means for the "old song," "mismatch," and new melody with "old words" conditions were 4.26, 3.88, and 3.05, respectively. With subjects as the sampling variable, the result of an analysis of variance was significant, $F(2,58)=28.18$, $p < .001$, and Newman-Keuls tests indicated that melody recognition under the "old song" condition was significantly better than that under the "mismatch" condition, $p < .05$.

With items as the sampling variable, an analysis of variance was performed on means generated only by the 18 subjects who had completed all items in all conditions. Those means, for the "old song," "mismatch," and new melody "old words" conditions respectively, were 4.10, 3.71, and 3.04. The main effect was significant, $F(2,34)=12.02$, $p < .001$, and *a priori* comparisons involving only the first two conditions revealed significance at the .02 level. (The results of *post hoc* tests were not significant, however.)

The results of the present experiment show that the integration effect is obtainable with phonetically similar nonsense used at test. One plausible explanation for this result is that words exert specific, albeit subtle, and memorable effects on the melodies with which they are sung. These submelodic effects include the manner in which consonants (perhaps also vowels) affect the onset, duration, and offset of particular melody tones. In our view, it seems indisputable that words exert variable effects on melody tones, as can be easily imagined, for example, in the case where two tones accompany the words "tip-top" as opposed to "ho-hum."

We think it not an accident that the present experiment showed evidence favorable to the submelodic hypothesis, whereas earlier efforts with a similar experimental design did not. The rules for deriving phonetically-similar nonsense texts were more fastidious here than those used before: For example, in these new materials we respected the voiced/voiceless distinction more consistently than under the old rules. Consonants with stop closures were distributed equally in the original and derived versions, too. These distinctions are just the sort that would be expected to underlie a submelodic effect of words on music.

Other interpretations of the integration effect, for example those to be considered below, might also be consistent with the evidence adduced here for the submelodic hypothesis. Comparing Experiments 2 and 3 of the present series, we note a smaller, and statistically weaker integration effect in the latter, with the derived nonsense words, than in the former, with the very same nonsense texts presented at learning and test. This is as it should be, by any commonsense view, for no scheme for deriving "similar" phonetic texts could possibly be as faithful a reinstatement as complete identity. On the other hand, we should not exaggerate the triumph of the submelodic hypothesis: At most, we can claim that we have shown conclusively that some such factor is operating somehow in our integration experiments, not that it is an answer as to the complete cause of the effect.

Introduction to Experiments 4 through 6

The remaining three experiments investigated the degree to which the melodies and texts might be associated in memory because of their close temporal proximity, as successive events in Experiments 4 and 5, and as simultaneous events in Experiment 6. These experiment address what we referred to above as the *association-by-contiguity hypothesis*. The term *association*, by itself, may connote many things theoretically, such as rote learning, Pavlovian conditioning, and pre-cognitive, antediluvian mists of antiquity. However, its denotation is theoretically empty: It simply stands for an experimental fact, that events A and B stand in a particular empirical relationship because of their history of co-occurrence. The challenge for theory is to rationalize the circumstances necessary for that association to be formed and the nature of the bonding thereby achieved. Thus, our integration result illustrates some form of association, without doubt. The submelodic mechanism, for which we adduced some support in Experiments 1 to 3, is not strictly an associative mechanism at all, but rather an effect of one element upon the nature of the other, namely that the occurrence of A with B changed the *physical nature* of B. We now ask whether the temporal contiguity of A and B, words and melody respectively, is a sufficient condition for their association when no possibility exists for an overt influence of one upon the physical integrity of the other, as with the submelodic mechanism.

In considering the theory of associations we have relied upon the distinction in the respective psychologies of James Mill and of John Stuart Mill between *mental compounding* and *mental chemistry* (see Boring, 1957, chapter 12). In the

former case two components retain their independent identities, yet are connected to one another. In the latter case the two components are themselves altered by each other's presence. Our concept of the *association-by-contiguity* of melody and text is like that of mental compounding: the melody and text are connected in memory, hence act as recall cues for each other, yet each is stored with its independent integrity intact. By contrast, the submelodic hypothesis in the first three experiments is consistent with a somewhat more chemical form of bonding, for a melody and text change each other physically when sung together in a song. A more purely *mental* chemistry could be an associative process in which, by co-occurrence *in the mind*, the memory representation of each is changed as against what it would have been without a particular companion.

More recently, a similar distinction has been articulated by Horowitz and Manelis (1972), albeit with a linguistic orientation, for adjective-noun phrases. They refer to a distinction between *I-Bonding* (where I stands for *individual* or *independent)* and *J-Bonding* (where J stands for *joint*). The former, illustrated by the phrases *deep-chair* or *dark-wing*, take their meaning as a phrase from the meanings of the constituent words. The latter, illustrated by *high-chair or right-wing*, possesses idiomatic meaning that transcends the meanings of the several constituents. As Horowitz and Manelis remarked (p. 222), I-Bonding owes allegiance to the British empiricist philosophers and J-Bonding to the Gestalt tradition. Tulving's work on recognition failure in episodic memory (Tulving, 1983) illustrates the same properties as J-Bonding, wherein an element of an association can be only poorly recognized but can be well recalled given the original associate as a cue. In many ways we believe that these issues are raised in their most stark relief when the two constituents, such as words and melodies, are fundamentally different cognitive elements than when intraverbal associations are at stake.

Experiment 4

The present experiment investigated the concept of contiguity as a sufficient condition for association. It assessed the degree to which a text could serve as the retrieval cue for a melody, when the two had initially been heard in close temporal proximity (in this case successively), yet not as a proper song. Each component in Experiment 4 was strictly independent physically: Texts were spoken and melodies were hummed. Using a technique similar to those reported above, we gave subjects a serial presentation of spoken texts, each followed by its corresponding melody, hummed, and then each text-melody pair was followed by a 10-sec interval of silence during which subjects were to "imagine" the song. A melody recognition test followed in which true (sung) songs and hummed melodies were heard. If subjects had managed to imagine the songs, as instructed during original presentation, they should have behaved in the same way as subjects in our earlier experiments, who had actually heard the songs. The test items were of five types (quotation marks indicate a deviation from what was heard in the presentation):

(a) "old songs" (the text and melody of one pair from the presentation were sung together as a song in testing);

(b) "mismatch songs" (the text from one pair and the melody from a different pair were sung together as a song);

(c) "new melody/old words" (the text from one pair heard in the presentation was sung with a new melody);

(d) old melody hummed (exactly as heard in the presentation except not preceded by words);

(e) new melody hummed.

The main question was whether the "old song" condition would generate better melody recognition than would the "mismatch" condition. Such an advantage could derive from either of two processes, corresponding to I- or J-Bonding in the terminology of Horowitz and Manelis (1972). If the simple contiguity hypothesis were correct, we would expect that melodies and texts presented in close succession would be connected in memory, hence could act as recall cues for one another. Thus, in a melody recognition test consisting of true (sung) songs, the melody should be better recognized when heard with the text with which it is connected than with a different (mismatched) but equally familiar text. Likewise, as we said above, if subjects are able to fuse the melodies and text mentally, using the 10 sec "imagine" period as they were instructed, to create a song-like memory representation, then, too, the old song condition would produce better melody recognition than the mismatch condition, as in the earlier experiments. Thus, the conditions of this experiment could not permit a choice between these two hypotheses. That would require consultation of still further experimental arrangements. However, a positive

outcome here would be a necessary, if insufficient, condition for either hypothesis.

Method

Materials

The 20 pairs of folksong excerpts were used with their real (not nonsense) words. Tape recordings of spoken texts and hummed melodies were made by the same female alto that was the singer in our earlier studies. Each spoken text generally followed the rhythm of its companion melody, consumed approximately the same amount of time, and had the character of poetic speech rather than normal, conversational speech.

Design

The design was comparable to that of Experiment 1, with five sets of tapes used to counterbalance the test conditions for each presentation item across five subject groups. The presentations consisted of 24 text-melody pairs, where the spoken text and hummed melody in each pair were separated by a one-sec interval and followed by 10 sec of silence. The test consisted of 20 items, four each assigned to the five conditions.

Procedure

The procedure was generally the same as that of the earlier experiments, except that subjects were told they would hear spoken texts and hummed melodies from simple folksongs and that they were to use the 10-sec interval to "imagine the words and melody together as though someone were singing them" and to "sing the song in your head." On the test, subjects were told to expect either true (sung) songs or hummed melodies and to indicate melody recognition as in the other experiments.

Subjects

Twenty five adults with undetermined levels of musical training were divided equally among the five groups.

Results and Discussion

As in earlier experiments, melody recognition ratings had a theoretical range of 1 to 6. Mean ratings respectively for the "old song," "mismatch song," "old words/new melody," old melody hummed, and new melody hummed were 4.24, 4.08, 3.63, 4.28, and 3.23. The result of an analysis of variance with subjects as the sampling variable gave a statistically significant main effect of condition, $F(4, 96) = 5.38$, $p < .001$. Newman-Keuls tests showed that melody recognition, on its own, was better than chance; there was a significant difference in the baseline comparison between the means for old and new hummed melodies (4.28 and 3.23 respectively), $p < .05$. This conclusion is bolstered by the fact that the melodies of "old songs" yielded a higher mean rating (4.24) than did "new melodies with old words" (3.63), $p < .05$. However, no integration effect for texts and melodies occurred. That is, the mean rating for "old songs" (4.24) was not significantly higher than that for "mismatch songs" (4.08). There was, then, no advantage for melody recognition conferred by the presence of original words over different but equally familiar words.

Thus, we can report no evidence that a melody and text heard in succession are better recognized later in each other's presence, even when instructions had been given to imagine the two presentation components as a song. Several reasons could possibly underlie the failure to find an association effect here. The problem may have been in the generation process, for one thing: Subjects may have been simply unable to imagine the combined components as a unified song. Or, the imagination instructions themselves may somehow have served to distract the subjects from encoding the two components even as "normal" paired associates. Further, in this experiment an unprecedented inconsistency existed between what was heard in input (successive spoken speech and hummed melodies) and what was tested for recognition in output (sung songs); this may have been distracting. And of course it is possible that spoken and hummed stimuli of the sort employed here are not conducive to *either* a process of song generation or of contiguous associative bonding.

A special circumstance introduced by spoken speech and hummed melodies is the introduction of elements in each stream that are foreign to the identity of these as components within normal songs: The prosody of normal speech necessarily introduces intonation gestures (phrasal declination for example) that would not be compatible with the sung version. The act of humming, likewise, cannot but introduce nasal and vowel segments that might otherwise not reside in the spoken text. Therefore, our abstraction of the spoken and melodic streams in the methodology of this set of studies is not an absolutely neutral operation. As so often, negative results are ambiguous, but positive results (see below) speak with considerably greater force.

In our next experiment we focused on the possibility that the instruction to generate songs, at presentation, had backfired even to the extent

of preventing the formation of independent (I-bonded) associations.

Experiment 5

Experiment 5 differed from Experiment 4 in two ways: First, no instructions for imagining songs were given at presentation. Second, test items contained the same format of successive components as had the presentation, that is, spoken words followed by hummed melodies. Abandoning the imagery instruction was intended to determine whether a melody and text could become independently associated in memory if the two components had been in close temporal proximity. Having established baseline melody recognition in Experiment 4, we saw no need to repeat the two hummed-melody conditions. The melody recognition test consisted of successive text/melody pairs under the following three conditions:

(a) old songs (a spoken text followed by its hummed melody, exactly as heard in the presentation); [3]
(b) mismatch songs (a spoken text followed by the hummed melody of a different text/melody pair from presentation);
(c) old text with a new melody that had not been heard in the presentation.

Method

Materials

Eighteen of the 20 song pairs used in Experiment 4 were used in the present experiment.

Design

The design was comparable to that of previous Experiments 2 and 3. Three sets of tapes counterbalanced the test conditions for each presentation item across three subject groups. The presentation consisted of 24 text/melody pairs, and the test consisted of 18 text/melody pairs, 6 each of the three conditions.

Procedure

The procedure was comparable to the earlier experiments. Subjects were told to expect pairs of spoken texts and hummed melodies on the presentation and test. Melody recognition ratings were obtained as before.

Subjects

Fifteen adults with undetermined levels of musical training were equally divided among the three groups.

Results and Discussion

As before, melody recognition ratings had a theoretical range of 1 to 6. Means for the old song, mismatch, and old words/new melody conditions were 4.22, 4.12, and 2.98 respectively. Clearly, no significant difference emerged between the old song and mismatch conditions. In other words, hummed melodies were not better recognized when preceded by the same spoken text which had preceded that melody at presentation than when preceded by a different (yet equally familiar) text. Thus, no evidence was found that a link in memory is engendered by the successive presentation of independent texts and melodies. This leaves open the possibility that mental compounding is not the agency for the integration effect between melody and text reported in our earlier experiments. If not, then the submelodic hypothesis remains the only explanation for the effect with evidence in its favor. However, a lingering question is whether a *simultaneous* presentation of spoken text and hummed melody could give rise to an association in memory of the two components. This was addressed in the following experiment.

Experiment 6

The objective of this experiment was to assess the degree to which a simultaneous presentation of spoken words and hummed melody could give rise to an integration of the two such that the melody was recognized better in the presence of the text with which it had originally been presented than in the presence of a different (equally familiar) text. This result would indicate that our reasoning about independent associative bonding had been correct, in Experiments 4 and 5, but our realization of contiguity had been inadequate.

The presentation episodes consisted of normal spoken texts and hummed melodies heard simultaneously and binaurally (but not dichotically). We refer to these simultaneous pairings as "spoken songs." The later recognition tests were of two types: Half the subjects heard only spoken songs (as in presentation) and half heard true, sung songs. Again, no instruction for the generation of song-like representations was given. So the question at hand was whether an association between contiguous components, if it occurred, would influence melody recognition *only* if the test stimuli were like those of the presentation or whether that association's influence would extend also to the case of true songs.

The melody recognition tests for both the (between-subject) conditions with spoken songs and true songs consisted of three within-subject conditions. As before, the critical comparison was that between old songs and mismatch songs.

(a) old songs (same text and melody as was heard in the presentation);
(b) mismatch songs (the text of one pair and the melody of a different pair heard in the presentation);
(c) old words with new melody.

Method

Materials

Eighteen of the 20 song pairs were used.

A master tape, from which experimental tapes were dubbed, was prepared by the same alto singer, as follows: Hummed melodies were first recorded in succession, each preceded by exactly four evenly spaced taps, also recorded onto the tape. The resulting signal was then fed into a second tape recorder at the same time that spoken texts were recorded onto a second tape. The singer listened to the hummed melodies from the first tape over headphones, using the four taps to fix the onset of the hummed melody, and then spoke the text along with the melody, recording both onto the second tape. Texts were generally spoken in the rhythm of the melody and also began and terminated in synchrony with it. When experimental tapes were dubbed from the master, the four taps were omitted. The test tapes employing true songs were the same as those employed previously with these materials.

Design

The design was exactly analogous to that of Experiment 2, except that two sets of test tapes were constructed, each administered to a different group of subjects, one set with spoken songs and the other with true songs.

Procedure

The procedure was comparable to that of the earlier experiments, with subjects told to expect the spoken texts of simple folksongs to be heard simultaneously with hummed melodies. At test, one group was told that items would be true, sung songs, and the other group that test items would be similar to presentation items. In all cases, of course, instructions called for recognition based only on the melodies.

Subjects

Twenty-four adults with undetermined levels of musical training were equally divided between the two test groups.

Results and Discussion

Melody recognition ratings had a theoretical range of from 1 to 6. Mean ratings for old songs, mismatch songs, and old words/new melody respectively were 4.56, 4.04, and 3.33 when the test consisted of spoken songs (as heard in the presentation) and 4.35, 3.96, and 3.25 when the test consisted of true songs. Two mixed analyses of variance were performed with type of test (spoken vs. true songs) as a between-subjects variable and the same three conditions ("old song," "mismatch," and new melody/"old words") as a within-subjects variable. With subjects as the sampling variable, only the main effect of conditions was significant, $F(2,44)=21.68$, $p < .001$; neither the type of test main effect nor the interaction was significant. The Newman-Keuls test indicated that combined "old song" ratings for both groups (4.46) exceeded that for "mismatch songs" (4.00), $p < .05$. Similarly, with items as the sampling variable, only differences among the three conditions were significant, $F(2,68)=13.65$, $p < .001$. The Newman-Keuls again test supported the difference between "old song" and "mismatch" ratings, $p < .05$.

Thus, by the reasoning of the fourth and fifth experiments here, true temporal contiguity was the necessary condition for observing our integration effect. The most straightforward interpretation of that result is that, in Experiment 6, conditions were favorable for the formation of independent associative links between constituents that had not lost their individual identity. Close temporal proximity, as in Experiments 4 and 5, was apparently not enough.

Here, for the first time in this series, we may rule out the submelodic hypothesis, because the pairing manipulation cannot have had any substantial effect on the physical nature of each constituent.[4] Likewise, the cognitive version of the submelodic hypothesis—J-Bonding indicative of what we have called "mental chemistry"—received no encouragement from these last three experiments. What makes a difference is not whether or not people try actively to integrate the melody and words in their minds, constructing an unheard song, but whether or not the two were strictly simultaneous.

Straightforward as this interpretation is, we are not so naive as to believe that one absent interaction from a single experiment can overthrow ideas as important as J-Bonding or Encoding Specificity. Besides the usual caution that we need more converging evidence on this point, it could be argued, albeit with some considerable added assumptions, that all subjects, in both groups of this experiment, left the presentation sequence with self-generated songs as memory representations. Hearing a hummed melody at the same time as one hears a rhythmically-matching stream of words might produce the experience of a song, whether the subject is trying to generate this or not. This would account for the integration effect among subjects tested with real songs. To account for the same effect in subjects tested with spoken songs, we need only observe that for these people, the conditions of acquisition and testing were exactly the same, which could have outweighed the disadvantage produced by the need for these subjects to generate song representations at test, as well as at acquisition.

An automatic process generating song-like representations from simultaneous, compatible verbal and musical streams would not be unexpected from a consideration of speech processing: In our ordinary lives, a simultaneous mixture of this sort is the rule rather than the exception, because the segmental features (in words) are always overlaid upon supra-segmental features, including specifically variation in fundamental frequency. For this ecological reason, simultaneous variation in pitch might be assigned to the prosodic aspect of speech automatically, even when the listener "knows" the verbal and tonal messages are nearly independent, as in listening to songs, or spoken songs. These considerations lead us to the design of future experiments better exploiting the tonal prosody and spoken message of integrated language communication.

General Discussion

As for the integration effect in song, our experiments in this and in the two previous papers (Serafine et al., 1984; Serafine et al., 1986) have guided our thinking in a number of ways. First of all, and despite musicological folklore to the contrary, the meaning of words seems to have a negligible role in the fact that melody and words of folksongs become stored in an integral fashion. Here again, in Experiment 2, the result withstood nonsense materials devoid of conventional meaning.

Secondly, we have adduced statistically reliable support for the submelodic hypothesis, suggesting that particular words can change the musical line sufficiently to influence recognition of the melodies later. It is no wonder people are slow to realize that "Baa, Baa, Black Sheep" and "Twinkle, Twinkle, Little Star" are words to the same tune—they are *not*, musically, quite the same tune, by virtue of the words to which each has been set.

Finally we have uncovered a number of factors that govern the size of the effect, some statistically reliable on their own and others not. In retrospect, it was perhaps misguided for us to have thought that a single factor would control the integration effect. Among the agencies for which evidence exists, we must include first temporal contiguity, as shown in Experiment 6 here. Barring the unknown contribution of automatic fusion of text and melody in songs, hearing words and the melody at the same time appears to affect their joint storage in the manner of paired associates. But we should not discard completely those factors uncovered by earlier experiments in this series as potential factors; even though they were not reliable on their own, they did measurably affect the size of the effect. Among these, we count instructions to attend only to melodies rather than to the whole songs at presentation (Serafine et al., 1984). Similarly, in the same experiment, acoustic non-identity of presentation and test materials (different singers, respectively) had an effect in the direction that would have been predicted (though not significantly). Elsewhere (Serafine et al., 1986 and here, in Experiment 4) we found that melodies in the presence of the *wrong* words did indeed have a distracting effect on melody recognition, beyond the facilitation that the correct words had.

Putting all these factors together, we believe we know well how to arrange conditions so as to maximize, or minimize, the integration of words and melodies in recognition of songs. This laboratory control is not unsatisfactory as explanation, provided one gives up the goal of having only one crucial component.

REFERENCES

Boring, E. G. (1957). *A history of experimental psychology*. New York: Appleton-Century-Crofts.

Erdei, P. (1974). *American folksongs to read, sing, and play*. New York: Boosey and Hawkes.

Horowitz, L. M., & Manelis, L. (1972). Toward a theory of redintegrative memory: Adjective-noun phrases. In G. H. Bower (Ed.), *The psychology of learning and motivation, 6* (pp. 195-224). New York: Academic Press.

Serafine, M. L., Crowder, R. G., & Repp, B. H. (1984). Integration of melody and text in memory for song. *Cognition, 16*, 285-303.

Serafine, M. L., Davidson, J., Crowder, R. G., & Repp, B. H. (1986). On the nature of melody-text integration in memory for songs. *Journal of Memory and Language*.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

FOOTNOTES

Memory & Cognition, in press, (in a shorter version).

†Yale University.

[1]We are well aware that the dictionary meaning of the word *contiguity* stipulates that the events in question be juxtaposed, or adjacent in time, but not overlapping or coterminous. This departs from usage of the term within psychology, where successive and simultaneous arrangements are both considered contiguous. In this paper we remain with this latter usage even though the former might be more justifiable to some scholars.

[2]Throughout this paper, quotation marks on test item labels indicate a deviation from the nomenclature described under General Method. Here, for example, an "old song" is so labelled because it is the real-word phonetic equivalent of an old song and is not exactly what was heard in the presentation.

[3]The stimuli were, of course, in no sense true *songs*. However, we retain the same terminology as used in the other experiments.

[4]Certainly not in Experiment 4 and 5, where the two constituents did not overlap in time. In Experiment 6, with simultaneous contiguity, masking-like effects could have existed between the melodies and texts. This perceptual interaction is not what we mean by physical interaction, which could not have occurred in any of these experiments.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 162-171

Orthography and Phonology: The Psychological Reality of Orthographic Depth*

Ram Frost†

The representation of meaning by words is the basis of the human linguistic ability. Spoken words have an underlying phonologic structure that is formed by combining a small set of phonemes. The purpose of alphabetic orthographies is to represent and convey these phonologic structures in a graphic form. Just as languages differ one from the other, orthographic systems represent the various languages' phonologies in different ways. This diversity has been a source of interest for both linguists and psychologists. However, while linguistic inquiry aims to explain and describe the origins and characteristics of different orthographies, psychological investigation aims to examine the possible effects of these characteristics on human performance. Consequently, reading research is often concerned with the question of what is universal in the reading process across diverse languages, and what aspects of reading are unique to each language's orthographic system. My first objective in this chapter is to outline the properties of different alphabetic systems that might affect visual word processing. The second objective is to provide some empirical evidence to support the claim that reading processes are determined in part by the language's orthography.

Orthography, phonology and the mental lexicon

The purpose of orthographies is to designate specific lexical candidates. There is, however, some disagreement as to how exactly this purpose is achieved. The major discussions revolve around the role of phonology in the process of visual word recognition. Clearly, phonologic knowledge of words generally precedes orthographic knowledge; we are able to recognize many spoken words long before we are able to read them. Only later, in the process of learning to read, does the beginning reader master an orthographic system, based, in western languages, on alphabetic principles.

The recognition of a printed word is based on a match between a letter string and a lexical representation. This match allows the reader access to the mental lexicon. However, since lexical access can theoretically be mediated by two types of abstract codes: orthographic and phonologic, a question remains about the exact transform of the printed word that is used in the process of visual word recognition: Is it informationally orthographic or phonologic?

One account argues that access to the mental lexicon is mainly phonologic (e.g., Liberman, Liberman, Mattingly, & Shankweiler, 1980). According to this view, orthographic information is typically recoded into phonologic information at a very early stage of print processing. Thus, the lexical access code for printed word perception is similar to that for spoken word perception. The appeal of this model is its parsimony and efficiency of storage; the reader does not need to build a visually coded grapheme-based lexicon, one that matches each of the words to spelling patterns in the language. Instead, a relatively small amount of information—knowledge of grapheme to phoneme correspondences—can recode print into a form every reader already knows: the speech-related phonologic form.

The second approach argues for the existence of an orthographic lexicon in addition to the phonologic one. According to this alternative view, lexical access for print can be achieved through

This work was supported in part by National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. Many ideas in the present chapter were generated in collaboration with Shlomo Bentin. I am also indebted to Laurie Feldman, Len Katz, and Ignatius Mattingly for their criticism on earlier drafts of this paper.

either system. The extreme position of this approach holds that lexical access is typically based only on the visual (orthographic) information, and the word's phonology is retrieved after lexical access has occurred. Possible exceptions are novel or low-frequency words that may lack an entry in the visually based lexicon (Seidenberg, Waters, & Barnes, 1984; Seidenberg, 1985). The appeal of such models is that visual lexical access is direct and, presumably, faster without the need for a mediating phonologic recoding. However, a model based on visual lexical representations must assume the existence of a memory store of orthographically coded words that parallels, in orthographic coding, most of the information the reader already possesses as phonologic knowledge.

Clearly, the reader is well aware of both orthographic and phonologic structures of a printed word. Hence, the debate concerning orthographic and phonologic coding is merely a debate about priority: is phonology necessary for printed word recognition to occur, or is it just an epiphenomenon that results from it? In other words: is phonology derived pre-lexically from the printed letters and serves as the reader's code for lexical search, or, rather, is lexical search based on the word's orthographic structure while phonology is derived post-lexically?

This question is often approached by monitoring and comparing subjects' responses in the lexical decision and the naming tasks. In lexical decision the subject is required to decide whether a letter string is a valid word or not, while in naming he is required to read the letter string aloud. In both tasks reaction times and error rates are measures of subjects' performance. Note that lexical decisions can be based on the recognition of either the orthographic or the phonologic structure of the printed word. In contrast, naming requires explicitly the retrieval of the printed word's phonology. Phonology, however, can be generated either pre-lexically by converting the letters into phonemes, or post-lexically by accessing the mental lexicon through the word's complete orthographic structure, and retrieving from the lexicon the phonologic information.

Since, at least theoretically, these two alternative processes are available to the reader, one should compare their relative efficiency. It has been suggested that the ability to rapidly generate pre-lexical phonology depends primarily on the reader's fluency, task characteristics, and the printed stimuli's complexity (see McCusker, Hillinger, and Bias (1981), for a review). In our present context, only the factor of stimulus complexity is of a special interest. Complexity is generally related to the amount of effort needed for decoding a given word. One possible source of complexity that merits close examination is the lack of transparent correspondence between orthographic and phonologic subunits. Because the purpose of orthographic systems is the representation of phonology, whether the skilled reader uses this information or not, the relative directness and simplicity—the transparency—of this representation can be of major importance.

Orthographic depth—Evidence from the shallow Serbo-Croatian

Although the transparency between spelling and phonology varies within orthographies, it varies more widely between orthographies. The source of this variance can be often attributed to morphological factors. In some languages, (e.g., in English), morphological variations are captured by phonologic variations. The orthography, however, was designed to preserve primarily morphologic information. Consequently, in many cases, similar spellings denote the same morpheme but different phonologic forms: the same letter can represent different phonemes when it is in different contexts, and the same phoneme can be represented by different letters. The words "heal" and "health", for example, are similarly spelled because they are morphologically related. However, since in this case, a morphologic derivation resulted in a phonologic variation, the cluster "ea" represents both the sounds [i] and [ ].

Within this context English is often compared to Serbo-Croatian. In Serbo-Croatian, (aside from minor changes in stress patterns), phonology almost never varies with morphologic derivations. Consequently, the orthography was designed to represents directly the surface phonology of the language: Each letter denotes only one phoneme, and each phoneme is represented by only one letter. Thus, alphabetic orthographies can be classified according to the transparency of their letter to phonology correspondence. This factor is usually referred to as "orthographic depth" (Klima, 1972; Liberman et al., 1980; Lukatela, Popadić, Ognjenović & Turvey, 1980, Katz & Feldman, 1981). An orthography that represents its phonology in an unequivocal manner is considered shallow, while in a deep orthography the relation of orthography to phonology is more opaque.

Katz and Feldman (1981) suggested that the kind of code that is used for lexical access depends

on the kind of alphabetic orthography facing the reader. Shallow orthographies can easily support a reading process that uses the language's surface phonology. On the other hand, in deep orthographies, the reader is encouraged to process printed words by referring to their morphology via their visual-orthographic structure. Note that orthographic depth does not necessarily have to have a clear psychological reality. For example, it has been argued that visual-orthographic access is faster and more direct than phonologic access (e.g., Baron & Strawson, 1976). By this argument, it might be the case that in all orthographies words can be accessed easily by recognizing their orthographic structures visually. Therefore, the relation between spelling and phonology should not necessarily affect subjects' performance.

Most of the earlier studies in word recognition were conducted with English materials. But in order to validate the psychological reality of orthographic depth experimenters turned to shallower orthographies like Serbo-Croatian.

In addition to its direct spelling to phonology correspondence, the Serbo-Croatian orthography has an additional important feature: It uses either the Cyrillic or the Roman letters, and the reader is equally familiar with both sets of characters. Most characters are unique to one alphabet or the other, but there are some characters that occur in both. Of these, some receive the same phonemic interpretation regardless of alphabet. These are called COMMON letters. Others receive a different interpretation in each alphabet. These are known as AMBIGUOUS letters. Letters string that include unique letters can be read in only one alphabet. Similarly, letters string composed exclusively of common letters can be read in only one way. By contrast, strings composed only of AMBIGUOUS and COMMON letters are bivalent. They can be read in one way by treating the characters as Roman graphemes and in distinctly different way by treating them as Cyrillic graphemes. The two alphabets are presented in Figure 1. This specific feature of the Serbo-Croatian orthography was used in several studies in order to examine phonological processing in visual word recognition (Lukatela et al. 1980; Feldman & Turvey, 1983)

Lukatela et al. (1980) investigated lexical decision performance in Serbo-Croatian, for words printed in the Cyrillic and the Roman alphabets. They demonstrated that words that could be read in two different ways were accepted more slowly as words than words that could be read in one way. Thus, the fact that one orthographic form had two phonologic interpretations slowed subjects' reaction times. This outcome suggested that the subjects were sensitive to the phonologic structure of the printed stimuli, while making lexical decisions. Lukatela et al. concluded that lexical decisions in Serbo-Croatian are necessarily based on the extraction of phonology from print. Similar results were found by Feldman and Turvey (1983) that compared phonologically ambiguous and phonologically unequivocal forms of the same lexical items. They have suggested that the direct correspondence of spelling to phonology in Serbo-Croatian results in an obligatory phonologic analysis of the printed word that determines lexical access. Moreover, in contrast to data obtained in English, the skilled reader of Serbo-Croatian demonstrates a bias towards a phonologically analytic strategy.

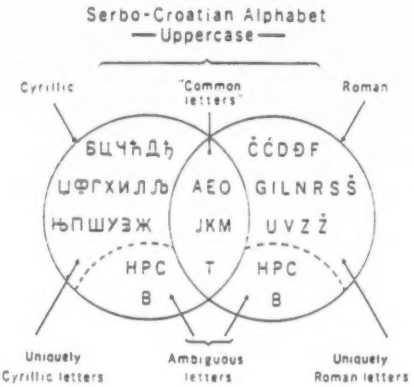


Figure 1.

Evidence from the deeper Hebrew orthography

The term "orthographic depth" has been used with a variety of related but different meanings. Frost, Katz, and Bentin (1987), suggested that it can be regarded as a continuum on which languages can be arrayed. They proposed that the Hebrew orthography could be positioned at the extreme end of this continuum, since it represents the phonology in an ambiguous manner.

Hebrew, like other Semitic languages, is based on word families that are derived from tri-consonant roots. Therefore, many words share an identical letter configuration. The orthography was designed primarily to convey to the reader the word's morphologic origin. Hence, the letters in Hebrew represent mainly consonants, while the vowels are conveyed by diacritical marks presented beneath the letters. The vowels marks,

however, are omitted from regular reading material, and can be found only in poetry, children literature or religious scripts (for a detailed description of the Hebrew orthography see Navon & Shimron, 1984). When the vowels are absent, a single printed consonantal string usually represents several different spoken words (sometimes up to seven or eight words can be represented by a single letter string). The Hebrew reader is, therefore, regularly exposed to both phonologic and semantic ambiguity. An illustration of the Hebrew ambiguous unvoweled print is presented in Figure 2.

Although it is clear that the Hebrew orthography is an example of a very deep orthography, this is for different reasons than those presented in the context of the English vs. Serbo-Croatian distinction. English is labeled as deep because of the opaque correspondence between single graphemes and phonemes in the language's spelling system. In contrast, this correspondence is fairly clear in Hebrew, since the consonants presented in print, aside from a few exceptions, correspond to only one phoneme. However, because the vowels are absent, the Hebrew orthography conveys less phonologic information than many other orthographies. Hence, it is not just ambiguous, it is incomplete. This characteristic of Hebrew, as I will argue, is not only linguistic but also psychological, in that it provides a possible explanation of differences in reading performance revealed in this language.

In order to assign a correct vowel configuration to the printed consonants to form a valid word, the reader of Hebrew has to draw upon his lexical knowledge. The choice among the possible lexical alternatives is usually based on contextual information: the semantic and syntactic contexts constrain the possible vowel interpretations. For an unvoweled word in isolation, however, the reader cannot rely on contextual information for the process of disambiguation.

Several studies have examined reading processes of isolated Hebrew words. Bentin, Bargai, and Katz (1984) examined naming and lexical decision for unvoweled consonantal strings. Some of these strings could be read as more than one word while some could be read as one word only. The results demonstrated that naming of phonologically ambiguous strings was slower than naming of unambiguous ones. In contrast, no effect of ambiguity was found in the lexical decision task. These results suggest that the reader is indeed sensitive to the phonologic structure of the orthographic string when naming is required. Contrarily, lexical decisions are not based on a detailed phonological analysis of the printed word in Hebrew. Note, that this outcome is in sharp contrast to the results obtained in the shallow Serbo-Croatian.

Figure 2. Unvoweled and voweled forms of the Hebrew tri-consonantal root ספר (sfr).

Lexical decisions and naming of isolated Hebrew words were further investigated in a study by Bentin and Frost (1987). In this study subjects were presented with phonemically and semantically ambiguous consonantal strings. Each of the ambiguous strings could have been read either as a high-frequency word or as a low-frequency word, pending upon different vowel assignments. Decision latency for the unvoweled consonantal string was compared to the latencies for both the high and the low-frequency voweled words. The results showed that lexical decisions for the unvoweled ambiguous strings were faster than lexical decisions for either of their voweled (therefore disambiguated) alternatives. This outcome was interpreted as evidence that lexical decisions for Hebrew unvoweled words were given *prior* to the process of phonological disambiguation. The decisions were probably based on the printed word's orthographic familiarity (cf. Balota & Chumbley, 1984; Chumbley & Balota, 1984). Thus, it is likely that lexical decisions in Hebrew involve neither a pre-lexical phonologic code, nor a post-lexical one. They are based upon the *abstract* linguistic representation that is common to several phonemic and semantic alternatives.

These results are in contrast to studies on lexical ambiguity conducted in English. Lexical disambiguation in English can be examined by employing homographs. Such studies have suggested that, at least initially, all meanings high- as well as low-frequency are automatically accessed in parallel. (Onifer & Swinney, 1981; Tanenhaus, Leiman & Seidenberg, 1979; and see Simpson, 1984, for a review). It should be noted, however, that in most cases the ambiguity in English resides only in the semantic and syntactic levels. With a few exceptions (e.g., "bow", "wind"), English homographs have only one phonologic representation, and the reader, usually, does not have to access two different words related to one printed form.

Although lexical decision in Hebrew might be based on an abstract orthographic representation, there is no doubt that the process of word identification continues until one of several phonological and semantic alternatives are finally accessed. This process of lexical disambiguation is more clearly revealed by using the naming task. Bentin and Frost (1987) investigated the process of selecting specific lexical candidates by examining the naming latencies of unvoweled and voweled words. In contrast to the result obtained for lexical decisions, naming of ambiguous strings was found to be just as fast as naming the most frequent voweled alternative, with the voweled low-frequency alternative slowest. In the absence of constraining context, the selection of one lexical candidate for naming seems to be affected by a frequency factor: the high-frequency alternative is selected first.

In a recent study (Frost & Bentin, in preparation), the processing of ambiguous consonantal strings in voweled and unvoweled Hebrew print was investigated by using a semantic priming paradigm. Subjects were presented with consonantal strings that could be read as a high- or a low-frequency word. These strings served as primes to targets that were related to one of the two alternative meanings. In order to minimize conscious attentional processes, targets followed the primes at a short SOA (stimulus onset asynchrony) of 100 ms (see Neely, 1976; 1977, for a discussion of this point). It was assumed that if a specific meaning of the ambiguous consonantal string was accessed, it would be reflected by a semantic facilitation for its respective target. Thus, lexical decisions for targets that are related to that specific meaning would be facilitated.

In contrast to studies on priming at short SOA's in English (e.g., Seidenberg, Tanenhaus, Leiman, and Bienkowsky (1982), no semantic facilitation for the low-frequency meanings was found in the unvoweled condition at 100 ms SOA. In the voweled condition there was a significant semantic facilitation for both the high- and the low-frequency meanings. This result suggests that in the voweled condition both the high-frequency and the low-frequency meanings of the consonantal strings were clearly depicted by the disambiguating vowel marks.

Apparently, since the Hebrew reader almost never reads voweled print, he uses the consonantal information for accessing the lexicon. The phonologic representation of the high-frequency is selected first. Only at a second stage does the reader consider the low-frequency alternative.

In conclusion, the deep unvoweled Hebrew orthography represents primarily the morphology of the Hebrew language, while phonemic information is conveyed only partially by print. Consequently, in addition to a phonologic lexicon the Hebrew reader has probably developed a lexical system which is based on phonologically and semantically abstract consonantal strings that are common to several words. Lexical processing occurs, at a first phase, at this

morphological level. The reader accesses the abstract string and recognizes it as a valid morphologic structure. Lexical decisions are usually given at this early stage and do not necessarily involve deeper phonological processing. The complete phonological structure of the printed word can only be retrieved post-lexically, after one word candidate has been accessed. The selection of a word candidate is usually constrained by context, but in its absence it is based on frequency factors.

Evidence from cross-language studies

Conducting experiments in different languages contributes important insights concerning the role of pre- or post-lexical phonology in deep and shallow orthographies. Nevertheless, conclusive inferences cannot be drawn from these studies unless they are supported by results obtained in cross-language designs. Cross-language designs allow a direct comparison of native speakers' performances when the independent variables under investigation are controlled between languages, under identical experimental conditions. Hence, they can provide direct evidence concerning the effects of the orthography's characteristics on the process of word recognition. Obviously, cross-language designs are not without potential pitfalls; language differences may be confounded with nonlinguistic factors. For example, differences in the subjects' samples due to motivation, education, etc., might interact with the experimental manipulation. The interpretation of the results, thus, hinges on whether they are likely to be free of such confounding.

Katz and Feldman (1983) compared semantic priming effects in naming and lexical decision in English and Serbo-Croatian. In this study, semantic facilitation was assumed to reflect lexical involvement in both tasks. The results demonstrated semantic facilitation for both lexical decision and naming in English. In contrast, semantic priming facilitated only lexical-decision in Serbo-Croatian. The authors suggested that phonology, which is necessary for naming, is derived post-lexically in English: hence the semantic facilitation in this task. In contrast, the extraction of phonology from print in Serbo-Croatian does not call for lexical involvement but is derived pre-lexically. An additional finding in the study was a high correlation of reaction times for lexical-decision and naming in Serbo-Croatian without semantic context. This result was interpreted as evidence for an articulatory code used in this language for both lexical decisions and naming.

The interpretation of differences in reading performance between two languages, as reflecting subjects' use of pre- vs. post-lexical phonology, can be criticized on methodological grounds. The correspondence between orthography and phonology is only one dimension on which two languages differ. English and Serbo-Croatian, for example differ in their grammatical structures, and in the size and organization of their lexicon (Lukatela, Gligorjević, Kostić & Turvey, 1980). These confounding factors, it can be argued, have affected subjects' performance in a similar way.

Frost, Katz, and Bentin (1987) endeavored to address this possible criticism by comparing three languages simultaneously. They examined lexical decision and naming performance in Hebrew, English, and Serbo-Croatian. Although any comparison between two of the languages might be confounded by other factors, the set of confounds is different for each of the three possible pairs of comparisons. The only factor that displays consistency with the dependent measure is orthographic depth. Assuming that it is indeed the main factor that influences subjects' performance, predictions concerning a two languages comparison should be extended to the third language. But, note that while the probability of obtaining a predicted correct ordering of performance in the two languages is one out of two, the probability is one out of six, when three languages are compared. Thus, an appropriate ordering of subjects' performance in three languages would corroborate more strongly the psychological reality of orthographic depth.

In their first experiment Frost et al. (1987) compared, in each language, reaction times for both lexical decision and naming of high-frequency words, low-frequency words, and nonwords, in English, Serbo-Croatian, and Hebrew. The results showed that the lexical status of the stimulus (being a high- or a low-frequency word, or a nonword), affected naming latencies in Hebrew more than in English, and in English more than in Serbo-Croatian. Moreover, only in Hebrew were the effects on naming very similar to the effects on lexical decision: Just as the lexical status of the stimulus affected lexical decisions, it also affected naming latencies. This outcome confirmed that in deep orthographies like Hebrew, phonology is derived post-lexically. In contrast, in a shallow orthography like Serbo-Croatian, naming performance is much less affected by lexical status. Given the direct

correspondence of orthography to phonology, the extraction of phonology from print does not call for lexical involvement.

In a second experiment, Frost et al. compared semantic priming effects in naming. Semantic priming usually facilitates lexical access. Hence, if the word's phonology is derived post-lexically in deep orthographies but pre-lexically in shallow orthographies, then naming should be facilitated more in Hebrew than in English, and again, more in English than in Serbo-Croatian. As hypothesized, the results revealed a relatively strong effect of semantic facilitation in Hebrew (21 ms), a smaller but significant effect in English (16 ms), and no facilitation in Serbo-Croatian whatsoever. These results were taken to strongly support the validity of the orthographic depth factor in word recognition.

In a recent study, Frost and Katz (1989) investigated how the different relations between spelling and phonology in English and Serbo-Croatian are reflected in the ability of subjects to match printed and spoken stimuli. They presented subjects simultaneously with words or nonwords in the visual and the auditory modality, and the subject's task was to judge whether the stimuli were the same or different. In order to carry out the matching process, the subjects had to mentally recode the print to phonology, and compare it to the phonologic information provided by the speech. Performance was measured in three experimental conditions: (1) Clear print and clear speech, (2) clear print and degraded speech, and (3) clear speech and degraded print. Within each language, the effects of visual and auditory degradation were measured relative to the baseline undegraded presentation.

When the visual or the auditory inputs are degraded, subjects are encouraged to restore the partial information in one modality by matching it to the clear information in the other modality. When subjects are presented with speech alone, restoration of degraded speech components has been shown to be an automatic lexical process (see Samuel, 1987). However, in addition to this ipsimodal restoration mechanism, subjects in the Frost and Katz experiment had the additional possibility of a compensatory exchange of speech and print information. Thus, the technique of visual and auditory simultaneous presentation and degradation provided insight concerning the interaction of orthography and phonology in the different languages.

The results showed that for Serbo-Croatian, visual degradation had a stable effect relative to the baseline condition (about 20 ms), regardless of stimulus frequency. For the English subjects, the effect of visual degradation was three to four times stronger than for the Serbo-Croatians. The inter-language differences that were found for visual degradation were almost identically replicated for auditory degradation: The degradation effects in English were again three to four times greater than in Serbo-Croatian. Thus, the overall pattern of results demonstrated that although the readers of English were efficient in matching print to speech under normal conditions, their efficiency deteriorated substantially under degraded conditions relative to readers of Serbo-Croatian.

These results were explained by an extension of an interactive model (see McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), that rationalizes the relationship between the orthographic and phonologic systems in terms of lateral connections between the systems at all of their levels. The structure of these lateral connections is determined by the relationship between spelling and phonology in the language: simple isomorphic connections between graphemes and phonemes in Serbo-Croatian, but more complex, many-to-one, connections in English. The concept of orthographic depth has direct bearing on the question of the relation between the phonologic and orthographic systems. Within such interactive models, the way in which connections are made between the two systems should be constrained by the depth of the orthography that is being modeled. In a shallow orthography, a graphemic node can be connected to only one phonemic node, and vice versa. Also, because words are spelled uniquely, each word node in the orthographic system must be connected to only one word node in the phonologic system. In contrast, in a deeper orthography, a graphemic node may be connected to several phonemic alternatives, a phonemic cluster may be connected to several orthographic clusters, and finally, a word in the phonologic system may be connected to more than one word in the orthographic system, as in the case of homophony (e.g., SAIL/ SALE) or, vice versa, as in the case of homography (e.g., WIND, READ, BOW, etc.). A representation of the different intersystem connections is demonstrated in Figure 3 for a word that exists in both the English and the Serbo-Croatian languages. The Serbo-Croatian word, KLOZET, is composed of unique letter-sound correspondences while the corresponding English word, CLOSET, is composed of graphemes, most of which have more than one possible phonologic representation, and phonemes,

most of which have more than one orthographic representation.

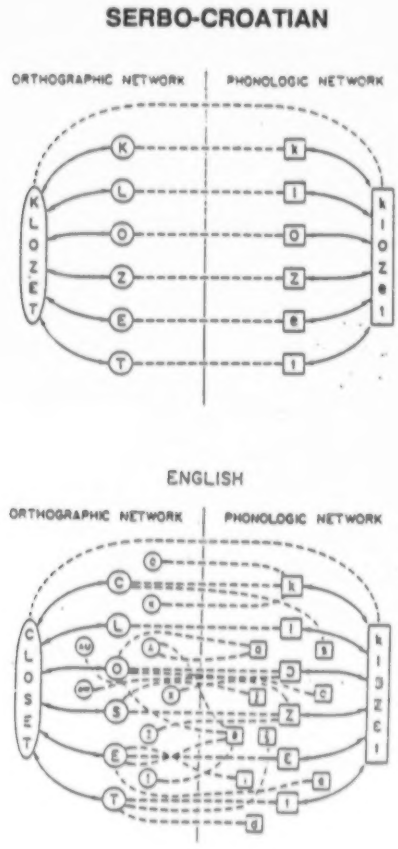


Figure 3.

As shown in Figure 3, the simple isomorphic connections between the orthographic and the phonologic systems in a shallow orthography should enable subjects to restore both the degraded phonemes from the print and the degraded graphemes from the phonemic information, with ease. This should be true because, in a shallow system, partial phonemic information can correspond to only one, or at worst, a few, graphemic alternatives, and vice versa. In contrast, in a deep orthography, because the degraded information in one system is usually consistent with several alternatives in the other system, the buildup of sufficient information for a unique solution to the matching judgment is delayed, and the matching between print and degraded speech, or between speech and degraded print, is slowed. Therefore, the effects of visual or auditory degradation was greater for English than for Serbo-Croatian.

The importance of orthographic depth: critique and conclusions

The psychological reality of orthographic depth is not unanimously accepted. Although it is generally agreed that the relation between spelling to phonology in different orthographies might affect reading processes to a certain extent, there is disagreement as to the relative importance of this factor. Seidenberg and his associates (Seidenberg et al. (1984); Seidenberg, 1985; Seidenberg & Vidanović, 1985) have argued that the primary factor determining whether or not phonology is generated prelexically is not orthographic depth, but word frequency. Their claim is that in *any* orthography, frequent words are very familiar as visual patterns. Therefore, these words can be easily recognized through a fast visually-based lexical access which occurs before a phonologic code has time to be generated pre-lexically from the print. For these words, phonologic information is eventually obtained, but only postlexically, from memory storage. According to this view, the relation of spelling to phonology should not affect recognition of frequent words. Since the orthographic structure is not converted into a phonologic structure by use of graphemes-to-phonemes conversion rules, the depth of the orthography does not play a role in the processing of these words. Orthographic depth exerts some influence, but only on the processing of low-frequency words and nonwords. Since such verbal stimuli are less familiar, their visual lexical access is slower, and their phonology has enough time to be generated prelexically.

In support of this hypothesis, Seidenberg (1985) demonstrated that there were few differences between Chinese and English subjects in naming frequent printed words. This outcome was interpreted to mean that in both logographic and alphabetic orthographies, the phonology of frequent words was derived postlexically, after the word had been recognized on a visual basis. Moreover, in another study, Seidenberg and Vidanović (1985) found similar semantic priming effects in naming frequent words in English and Serbo-Croatian, suggesting again that the phonology of frequent words is derived postlexically, whatever the depth of the orthography. These results are consistent with a recent study by Carello, Lukatela, and Turvey (1988), that demonstrated associative priming effects for naming in Serbo-Croatian. Although Carello et al. did not manipulate word-frequency in their study, their results question the inevitability of pre-lexical phonology in a shallow

orthography; some lexical influence on word recognition may be possible.

The resolution of these conflicting results is certainly not a simple task. A possible approach for examining the source of these differences could consist of examining the experimental characteristics of these studies. One salient feature of most of the experiments discussed above is that they were conducted exclusively in the visual modality; that is, print alone was used to study the relationship between orthography and phonology. The experimental manipulation of phonology, therefore, has been indirect, having been derived from manipulating the orthography. One can criticize this methodology for studying the processing consequences of the relation between phonology and orthography: Because phonologic variation is typically obtained through orthographic variation, one can never be certain which of the two is controlling the subject's responses. A simple example can be given in the case of homophones. The common assumption that two homophones (e.g., bear/bare; sale/sail), share a phonologic but not an orthographic structure (see for example, Rubenstein et al., 1971) is, in a way, misleading. Homophones always share printed consonants or vowels, and the task of disentangling the effect of the shared phonology from the shared orthography is complicated. Moreover, doubts have been raised about the adequacy of the lexical decision and naming tasks for measuring lexical as contrasted with prelexical involvement (see Balota & Chumbley, 1984, 1985).

The technique of simultaneous visual and auditory presentation with degradation proposed by Frost and Katz (1989); (see also Frost, Repp, & Katz, 1988), furnishes partial solutions to these methodological problems. First, phonology is presented to the subjects through a spoken word and does not have to be inferred from print. More importantly, by degrading the print or the speech, the technique affords a way to independently manipulate the perception of orthography and phonology. By using this method, Frost and Katz (1989) have demonstrated that orthographic depth and not word frequency is the primary factor that affects the generation of pre- or post-lexical phonology.

However the assessment of the role of orthographic depth in reading cannot be resolved solely with methodological arguments. One important conclusion from two decades of studies in reading is that the reader uses various strategies in processing printed words. (see McCusker et al., 1981). These strategies have been shown to depend on factors like orthographic regularity (Parkin, 1982), word frequency (Scarborough, Cortese, & Scarborough, 1977), ratio of words and nonwords (Frost et al., 1987), or special demand characteristics of the experimental task (e.g., Spoehr, 1978). By the same argument, one cannot fully account for the reader's processing without taking into consideration the reader's linguistic environment. Although the skilled reader in every orthography becomes familiar with his own language's orthographic structures, I suggest that the depth of the orthography is an important factor.

One common misinterpretation of claims concerning the importance of orthographic depth is to view a language's orthographic system as constraining the reader to only one form of processing. For example, although Frost et al. (1987) have shown no semantic facilitation for naming a specific set of stimuli in Serbo-Croatian, it does not follow that Serbo-Croatian readers never generate phonology post-lexically. One should always give the reader credit for extensive flexibility. If the words in the experiments were closely associated, even the Serbo-Croatian reader might find the extraction of phonology post-lexically more efficient then a pre-lexical extraction. But under similar conditions, relative differences should be found between deep and shallow orthographies.

In conclusion, the argument concerning the effect of orthographic depth is an argument concerning the *priority* of using a specific processing strategy for generating phonology in different orthographies. Research conducted in English Serbo-Croatian and Hebrew suggests that orthographic depth has indeed a strong psychological reality.

REFERENCES

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 340-357.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language, 24,* 84-106.

Baron, J., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 386-393.

Bentin, S., Bargai, N., & Katz, L. (1984). Orthographic and phonemic coding for lexical access: Evidence from Hebrew. *Journal of Experimental Psychology: Learning Memory & Cognition, 10,* 353-368.

Bentin, S., & Frost, R. (1987). Processing lexical ambiguity and visual word recognition in a deep orthography. *Memory & Cognition, 25,* 13-23.

Carello, C., Lukatela, G., & Turvey, M. T. (1988). Rapid naming is affected by association but not by syntax. *Memory & Cognition, 16*, 187-195.

Feldman, L. B., & Turvey, M. T. (1983). Visual word recognition in Serbo-Croatian is phonologically analytic. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 288-298.

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition, 17*, 302-310.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 104-114.

Frost, R., Repp, B. H., & Katz, L. (1988). Can speech perception be influenced by a simultaneous presentation of print? *Journal of Memory and Language, 27*, 741-755.

Katz, L., & Feldman, L. B. (1981). Linguistic coding in word recognition: Comparisons between a deep and a shallow orthography. In A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive processes in reading* (pp. 85-105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Katz, L., & Feldman, L. B. (1983) Relation between pronunciation and recognition of printed words in deep and shallow orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 157-166.

Klima, E. S. (1972). How alphabets might reflect language. In F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye*. Cambridge, MA: The MIT Press.

Koriat, A. (1985). Lexical access for low and high frequency words in Hebrew. *Memory & Cognition, 13*, 37-44.

Liberman, I. Y., Liberman, A. M., Mattingly, I. G., & Shankweiler, D. (1980). Orthography and the beginning reader. In J. F. Kavanagh & R. L. Venezky (Eds.), *Orthography, reading, and dyslexia* (pp. 137-153). Austin, TX: Pro-Ed.

Lukatela, G., Popadić, D., Ognenović, P., & Turvey, M. T. (1980). Lexical decision in a phonologically shallow orthography. *Memory & Cognition, 8*, 415-423.

McClelland, J. L., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

McCusker, L. X., Hillinger, M. L., & Bias, R. G. (1981). phonologic recoding and reading. *Psychological Bulletin, 89*, 217-245.

Navon, D., & Shimron, Y. (1984). Reading Hebrew: How necessary is the graphemic representation of vowels? In Leslie Henderson (Ed.), *Orthographies and reading: Perspectives from cognitive psychology, neuropsychology, and linguistics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition, 4*, 648-654.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General, 106*, 226-254.

Onifer, W. & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition, 9*, 225-236.

Parkin, A. J. (1982). phonologic recoding in lexical decision: Effects of spelling to sound regularity depends on how regularity is defined. *Memory & Cognition, 10*, 43-53.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 10*, 645-657.

Rumelhart, D., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*, 60-94.

Samuel, A. G. (1987). Lexical uniqueness effects on phonemic restoration. *Journal of Memory and Language. 26*, 36-57.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and Performance, 3*, 1-17.

Seidenberg, M. S. (1985). The time course of phonologic activation in two writing systems. *Cognition, 19*, 1-30.

Seidenberg, M. S.,Tanenhaus, M. K., Leiman, J. M., & Bienkowsky, M. (1982). Automatic access of the meaning of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology, 14*, 489-537.

Seidenberg, M. S. & Vidanović, S. (1985). Word recognition in Serbo-Croatian and English: Do they differ? Paper presented at the Twenty-fifth Annual Meeting of the Psychonomic Society, Boston.

Seidenberg, M. S., Waters, G. S., & Barnes, M. A (1984) When does irregular spelling of pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23*, 383-404.

Spoehr, K. T. (1978). Phonological recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 17*, 127-141.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior, 20*, 120-136.

Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence of multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18*, 427-440.

FOOTNOTES

*To appear in M. Noonan, P. Downing, & S. Lima (Eds.), *Literacy and Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Co.

†Now at the Department of Psychology, Hebrew University, Jerusalem, Israel.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 172-179

Phonology and Reading: Evidence from Profoundly Deaf Readers*

Vicki L. Hanson†

The prelingually, profoundly hearing-impaired reader of English is at an immediate disadvantage in that he or she must read an orthography that was designed to represent the phonological structure of English. Can the deaf reader become aware of this structure in the absence of significant auditory input? Evidence from studies with deaf college students will be considered. These studies indicate that successful deaf readers do appreciate the phonological structure of words and that they exploit this knowledge in reading. The finding of phonological processing by these deaf readers makes a strong case for the importance of phonological sensitivity in the acquisition of skilled reading, whether in hearing readers or deaf readers.

In the normal course of events, children are fluent speakers and listeners of their native tongue, English, before they begin learning to read. In this chapter, I am concerned with a population of readers for whom this is not the case. This population is prelingually, profoundly hearing-impaired. For these deaf readers, speech and lipreading are difficult to acquire and require years of instruction.

Research on reading has indicated that building on a spoken language foundation is a critical feature of reading and that in order to use an alphabetic orthography, such as English, to best advantage, the reader must go beyond the visual shape of words to apprehend their internal phonological structures (Liberman, 1971, 1973). Despite their extensive experience in using the phonology in everyday speech, evidence presented elsewhere in this monograph argues that hearing children who are poor readers may have phonological deficits that underlie their reading problem. These children have difficulty in setting up phonological structures, in apprehending such structures in words, and in using a phonetic code for the storage and processing of words in working memory. The phonological deficits of these children may be fairly subtle, however, such that no difficulty in the child's speaking ability or listening comprehension may be readily apparent.

If in the hearing population even subtle phonological deficits are associated with poor reading, then how is it possible for profoundly deaf individuals to read? One might suppose that deaf persons would have difficulty with reading, and, indeed, this is the case. Surveys have consistently shown that hearing-impaired students lag significantly behind their normally-hearing counterparts in reading achievement (Conrad, 1979; Furth, 1966; Trybus & Karchmer, 1977). Although it is typical to state, based on these surveys, that the average hearing-impaired student graduating from high school reads only at about the level of a hearing child of fifth grade, that statistic obscures the even greater reading deficiency of *profoundly* hearing-impaired students, that is, those who could be considered truly deaf. For them, the statistics are even more discouraging: Profoundly deaf students graduating from high school read, on the average, only at the level of a normally hearing child of third grade (Conrad, 1979; Karchmer, Milone, & Wolk, 1979). Remember, though, that these reading achievement scores

This writing of this paper was supported, in part, by Grant NS-18010 from the National Institute of Neurological and Communicative Disorders and Stroke. I wish to thank Carol Padden for her thoughtful reading of an earlier version of this manuscript.

represent only a population average. We find, for example, that measures of reading achievement for deaf students attending Gallaudet University may average seventh to tenth grade, with some students reading at above twelfth grade level (see, for example, Hanson, 1988; Hanson & Feldman, 1989; Hanson, Shankweiler & Fischer, 1983; Reynolds, 1975).

These statistics on reading achievement levels of deaf students have been used by investigators to argue two opposing views of the relationship between phonology and reading. The assumption common to both views is that the hearing impairment of these students prevents access to English phonology. In one view, access to phonological information is believed to be crucial for reading and the generally low reading achievement levels of deaf students are believed to reflect its importance. Because these readers presumably lack access to English phonology, their acquisition of reading suffers as a consequence. The second view takes the position that access to phonological information is not important in reading. The fact that *some* deaf individuals are able to attain fairly high reading levels is taken as evidence of this. Again, the assumption is that these readers, due to their hearing impairment, lack access to phonological information. Consequently, if they succeed at reading it must be without benefit of phonology.

Neither of these positions need be correct, however, in their interpretation of the reading achievement of deaf students. Deaf readers, despite their hearing impairment, might have access to phonology that could be used to support skilled reading. To assume that deaf readers lack access to phonology because of their deafness confuses a sensory deficit with a cognitive one. While the term *phonological* is often used to mean acoustic/auditory, or sound this usage reflects a common misunderstanding of the term. Phonological units of a language are not sounds, but rather a set of meaningless primitives out of which meaningful units are formed. These primitives are related to gestures articulated by the vocal tract of the speaker (see Liberman & Mattingly, 1985 for a more detailed discussion).[1] In the case of English, the deaf individual could learn about the phonology of the language from the motor events involved in speech production, through experience in lipreading, or from experience with the orthography.

As a rule, deaf children in English-speaking countries receive intensive instruction in speaking and lipreading. This is true both in schools that use an oral educational approach (with speech being the only means of communication used in the classroom) and in schools that use a simultaneous or total communication approach (with speech being accompanied by manual communication in the classroom). Through this speech training, prelingually, profoundly deaf individuals develop varying skill in speaking and lipreading. Although some of these individuals develop quite good speaking and lipreading skills, most do not (Conrad, 1979; Smith, 1975). Speech training, nevertheless, does provide the deaf individual with a means of learning English phonology.

Speech intelligibility does not necessarily indicate, however, the extent to which a deaf reader has access to phonological information. Intelligibility reflects the degree to which a deaf speaker's speech can be understood by a listener. Among the things that can effect intelligibility are phonation and prosodic information. While such features clearly add to intelligibility, they may not be relevant for an individual's internal manipulation of phonological information. In any event, it cannot simply be assumed that deafness necessarily blocks access to English phonology. This is a question for empirical investigation.

How do congenitally, profoundly deaf readers who read well manage to do it? That is the question to be addressed here. It is possible that deaf readers read English as if it were a logographic language; namely, treating printed English words as visual characters, without taking into account the correspondences between the printed letters and the phonological structure of words. Research on the reading of Japanese and Chinese, however, has suggested that for logographic languages, as for alphabetic languages, phonetic recoding of words is one component of a linguistic processing system required for the task of reading (Erickson, Mattingly, & Turvey, 1977; Mann, 1985; Tzeng, Hung, & Wang, 1977). For example, Tzeng et al. (1977) found that the phonetic composition of printed Chinese characters influenced sentence processing for skilled readers of Chinese. These investigators concluded that even in cases where lexical access is possible without phonological mediation, a phonetic code is still required for effective processing in working memory.

The deaf individuals who participated in the studies to be discussed here had backgrounds in which sign was used predominantly. That is, they generally *had* or *were* receiving instruction using sign language. Most of these individuals

considered American Sign Language (ASL), to be their preferred means of communication. ASL is the common form of communication used by members of deaf communities across the United States and parts of Canada. It is a visual-gestural language that has developed independently from spoken languages and from other signed languages. For many of the subjects in the studies reported here, ASL was their first language, having been learned as a native language from deaf parents. These subjects were typically undergraduates at Gallaudet University. All were profoundly deaf. These deaf subjects, therefore, can be characterized as having higher than average reading levels and not being exposed to an exclusively oral background.

Findings on phonetic coding in working memory

Evidence reviewed elsewhere in this monograph indicates that hearing children who are poor readers have a language deficit that is specific to the phonological domain. For example, in tests of short-term memory hearing poor readers recall fewer items overall and display less sensitivity to rhyme than hearing good readers (see, for example, Shankweiler, Liberman, Mark, Fowler, & Fischer, 1979). That is, on rhyming lists, the accuracy of the good readers is typically worse than on nonrhyming lists. In contrast, the accuracy of the poor readers is about the same for rhyming and nonrhyming lists. The good readers' differential performance on recall of rhyming and nonrhyming strings has been taken to mean that these readers convert the printed letters into a phonetic form and retain this phonetic information in memory. Accordingly, the finding that poor readers are not much affected by rhyming manipulations suggests that they are less able to use the phonetic information.

Is a phonetic code uniquely well-suited to the task of reading? To examine this question, we asked whether for deaf signers a different language code, one based on the structure of signs, could provide an alternative coding system for reading.

A sign of ASL is produced by a combination of the formational parameters of handshape, place of articulation, movement, and orientation (Battison, 1978; Stokoe, Casterline, & Croneberg, 1965). Evidence indicates that when *signs* are presented for recall in a short-term memory task, the signs are coded in terms of these formational parameters. The first line of evidence comes from studies of intrusion errors in the recall of lists of signs (Bellugi, Klima, & Siple, 1975; Krakow & Hanson, 1985). In the study by Bellugi, Klima, and Siple, lists of spoken words or signs were presented to hearing adults and deaf adults, respectively. The subjects were asked for immediate written recall. Intrusion errors for the hearing group were confusions of phonetically similar words. For example, a subject might write the word *boat* instead of the word presented, "vote." Errors for the deaf subjects, however, were completely different; they were confusions of the formational parameters of signs. As an example, a subject might write the word *egg* instead of the sign presented, NAME. The signs corresponding to the words *name* and *egg* differ only in terms of the movement of the hands (see Figure 1).

Figure 1. Formationally similar signs. Shown left to right from the top are KNIFE, EGG, NAME, PLUG, TRAIN, CHAIR, TENT, SALT. (From Hanson, 1982, p. 574).

In addition, there is evidence that lists of formationally similar signs can produce performance decrements in serial recall tasks (Hanson, 1982; Poizner, Bellugi, & Tweney, 1981). For example, shown in Figure 1 is a set of formationally similar signs that I used in one such study (Hanson, 1982). Shown here are, left to right from the top, the signs for the words KNIFE, EGG, NAME, PLUG, TRAIN, CHAIR, TENT, and SALT. On each trial in that experiment, deaf college students were shown five of the signs from this set, and were asked to remember the five signs in order. Results of that experiment indicated that fewer signs were recalled from lists made up of signs from this formationally similar set than from lists made up of signs from a formationally unrelated (control) set.

Despite such evidence that sign coding can thus mediate short-term recall of *signs*, evidence from other research does not support the notion that a sign code can serve as a viable code in the service of skilled adult reading. In another condition of that study (Hanson, 1982), I tested deaf college students in a short-term recall task of *printed* words. There were three types of word lists of interest here: rhyming words, orthographically similar words, and words whose signs were formationally similar. The words in the rhyming set were *two, blue, who, chew, shoe, through, jew,* and *you*. The words in the orthographically similar set were visually similar. The words in this set were *bear, meat, head, year, learn, peace, break,* and *dream*. While argument could be taken with the degree of visual similarity of the words in this list, it is at least true that these words are more similar visually than were the words in the rhyming list. The words in this visually similar set served as a control to ensure that any potential rhyme effects could not be attributed to the visual similarity of the printed words. The words in the formationally similar set were words whose corresponding signs were formationally similar. These were the words *knife, name, plug, train, chair, tent,* and *salt,* whose corresponding signs are shown in Figure 1. Each of these sets was paired with a control set of words. Of interest in this experiment was any differences in ability to recall an experimental and control set.

The pattern of results in that experiment clearly indicates the use of phonetic coding by the deaf subjects. Whereas these subjects recalled 65.4 percent of the lists in the control condition, they recalled only 47.6 percent of the lists in the phonetically similar (rhyming) condition. There was, however, no decrement on the visually similar lists, indicating that the decrement on the rhyming lists was due to phonetic, not visual, similarity.

Interestingly, I found no evidence that the deaf college students I tested were using sign coding. Their performance on lists of words having formationally similar signs and on the control lists was comparable (52.9 percent of the control lists recalled vs. 51.4 percent of the formationally similar lists recalled). Converging evidence from later research supports this finding that the better deaf readers do not use sign coding in their recall or reading of printed English words (Lichtenstein, 1985; Treiman & Hirsh-Pasek, 1983). More than 100 years ago, Burnet (1854) argued that sign coding would be a ponderous strategy for the deaf readers, and, thus, limited in its use to the poorer readers. By finding that the better deaf readers do not use sign coding when processing printed English words, current research on the cognitive processing of deaf readers is consistent with Burnet's speculations.

The finding that the better readers were using phonetic coding is reminiscent of the results reported by R. Conrad (1979) in a very large scale study of deaf and hearing-impaired students in England and Wales. Conrad tested these students in a short-term memory task of rhyming and nonrhyming lists of printed letters. Comparing their performance on this memory task with measured reading ability, Conrad found that the better readers in his deaf population recalled fewer rhyming than nonrhyming lists. Thus, the better readers were using phonetic coding.

My study with deaf signers (Hanson, 1982) took Conrad's findings one step further. Conrad's subjects were from schools that generally subscribed to an oral philosophy of education. As a result, phonetic coding was the only language form available to the subjects. In my study, the deaf subjects had sign language readily available to them. In fact, all of my deaf subjects had deaf parents and reported ASL to be their first language. Yet, these signers, as skilled deaf readers, used phonetic coding in that memory task, indicating the importance of phonetic coding in short-term retention of printed material.

Sensitivity to the phonological structure of English words

Additional evidence that deaf readers can access phonological information about English words is provided by studies of individual word reading. For example, one experimental paradigm that has been shown to produce phonological effects with

hearing readers uses a lexical decision task in which two letter strings are shown to the subjects on every trial, one string above the other (Meyer, Schvaneveldt, & Ruddy, 1974). The subjects must decide whether or not *both* of the letter strings on a trial are real English words.

In a series of three experiments, we used this paradigm with deaf college students (Hanson & Fowler, 1987). There were two types of word pairs of particular interest. As shown in Table 1, the first was pairs in which the two words rhymed. These rhyming words were spelled alike except for the first letter. The second type of word pair of interest was pairs in which the two words were spelled alike except for the first letter, but the pairs did *not* rhyme. It is apparent that the rhyming and nonrhyming pairs were equally similar orthographically, differing only in the phonological similarity of the two members of a pair. We tested whether there was any difference in the response times to the rhyming and nonrhyming pairs. Since, however, response times to words vary with word familiarity and orthographic regularity, it was not possible in this study to simply compare the responses to the rhyming and nonrhyming pairs. To eliminate familiarity and regularity as confounding factors, two control conditions were used. Word pairs in the control conditions used the same words as in the rhyming and nonrhyming pairs, but were repairings of these words. Thus, the control pairs for the rhyme condition were the same words as in the rhyme condition, just paired now with different words. For example, in the rhyme condition, the words *save-wave* and *fast-past* were paired together, while in the rhyme control *save-past* were paired together and *fast-wave* were paired together. Similarly, the control pairs for the nonrhyme condition were the same words as in the nonrhyme condition, just paired with different words. By comparing each word in the rhyme and nonrhyme condition with itself in a control condition, any effects of word frequency and regularity were eliminated.

Table 1. *Rhyming and nonrhyming pairs and their matched controls.*

Rhyming Pairs	Nonrhyming Pairs
save-wave	*have-cave*
fast-past	*last-east*
Rhyming Controls	**Nonrhyming Controls**
save-past	*have-east*
fast-wave	*last-cave*

Source of data: From Hanson & Fowler, 1987, Experiment 2.

The predictions for this experiment are shown in Table 2. If readers in this task did not access phonological information, then there should have been no effect due to the phonological relationships between words in a pair. That is, if the readers were using solely orthographic information, then the first equation shown here would hold; namely, that the difference between response times to rhyming pairs and the rhyme controls would equal the difference between response times to nonrhyming pairs and their controls. Thus, response times would be the same whether or not the words of a pair rhymed.

Table 2. *Predictions in the Hanson & Fowler, 1987 Study.*

If ORTHOGRAPHIC CODING, then:
Control - Rhyming = Control - Nonrhyming
If PHONOLOGICAL CODING, then:
Control - Rhyming ≠ Control - Nonrhyming

If, however, readers *were* accessing phonological information, then there *would* be a difference in response times as a function of phonological relationships between words in a pair. Access to phonological information would be indicated if the second equation held; namely, that the two differences in response times would not be equal. In that event, the response times would be affected by the rhyming manipulation.

For the deaf college students we tested, there *was* an effect of the phonological relationship between the words in a pair. Shown in Table 3 are the response times from one experiment of that study (Experiment 2). Response times were faster for the rhyming pairs than for the matched controls. In contrast, response times were slower on the nonrhyming pairs than on the matched controls. Since the rhyming and nonrhyming pairs were equally similar orthographically, this significant difference in response times for the rhyming and nonrhyming pairs was not due to orthographic influences. As a consequence, the difference in response times to the rhyming and nonrhyming pairs could be unambiguously attributed to the discrepant phonological structures. Thus, these good deaf readers, like hearing readers, accessed phonological information when reading words.

Most impressive is the finding that the deaf subjects in that study were not only accessing phonological information, but were doing so in a highly speeded task. It might be supposed that

deaf readers would be able to access phonological information only in situations in which they have time to laboriously recover learned pronunciations. In this research, however, we found that they accessed phonological information quite rapidly, suggesting that accessing such information is a fundamental property of reading for these skilled readers.

Table 3. *The response time (RT) difference for the rhyming and nonrhyming pairs and their matched controls for deaf college students.*

Control - Rhyming (52 ms) ≠ Control - Nonrhyming (-15 ms)

Source of data: From Hanson Fowler, 1987, Experiment 2.)

In more recent work, we have found other evidence that skilled deaf readers are sensitive to the phonological structure of words. For example, deaf college students, when asked to think of words that rhyme with a specific target word have been found to be able to do so (Hanson & McGarr, 1989). In addition, we have found that deaf college students are able to apply principles of grapheme-phoneme correspondence in generating the correct pronunciation of letter strings not previously encountered—a skill underlying the acquisition of new words. In this latter task (Hanson, 1989), I tested these students on their reading of orthographically possible nonwords; that is, pseudowords. The critical test was between pseudowords such as *flaim* that were homophonous with an actual English word (*flame*) and control pseudowords. These controls were orthographically-matched pseudowords that were not homophonous with an actual English word (e.g., *proom*). Examples of stimuli from the task are shown in Table 4.

Table 4. *Examples of pseudohomophones and control pseudowords.*

English Word	Pseudohomophone	Control
flame	flaim	proom
dog	daug	grine
spoon	spune	fosh
tall	taul	brate
home	hoam	spail
blue	bloo	nole
noon	nune	fune

Source: Selection of stimulus items from Hanson, 1989, based on Macdonald, 1988.

In two experiments using different lists of pseudowords, a paper and pencil task tested whether subjects could identify which of several pseudowords were homophonous with English words. The actual instructions to subjects were that they were to indicate whether or not each of the "nonsense words" was pronounced like a real English word. In both experiments, deaf college students were able to correctly make this judgment with better than chance accuracy, although they were not as accurate as the hearing subjects. As an additional aspect of this pseudohomophone task, subjects in one of the two experiments were asked to indicate *which* English word they thought a pseudoword was pronounced like, if they had indicated that they thought it was pronounced like one. In this second task, deaf subjects were usually able to supply the correct English word.

Studies on individual word reading thus indicate that it is possible for deaf readers to have access to English phonology. This does not mean that such access is easy for these readers. Nor does it mean that all or even most deaf readers are able to use this information. The point, rather, is that hearing loss alone does not preclude access to phonology. In addition, it is important to note and that the better deaf readers generally take advantage of this phonological information.

Why phonological coding?

In sum, the evidence, which has been summarized here, indicates that it is possible for deaf readers to use phonology. The use of phonological information tends to be characteristic of deaf good readers, whether they are beginning readers (Hanson, Liberman, & Shankweiler, 1984), high school students (Conrad, 1979; McDermott, 1984), or college students (Hanson, 1982; Hanson & Fowler, 1987; Lichtenstein, 1985). Why would the better deaf readers use this type of linguistic information when reading? One possibility has to do with the structural properties of particular languages. In English, where word order is relatively fixed, grammatical structuring is essentially sequential. A phonological code may be an efficient medium for retaining the sequential information that is represented in English.

Deaf individuals have specific difficulty in the recall of temporally sequential information (Hanson, in press). Studies have consistently found that the measured memory span of deaf individuals is shorter than that of hearing persons

(see, for example, Bellugi et al., 1975; Blair, 1957; Belmont & Karchmer, 1978; Conrad, 1979; Hanson, 1982; Kyle, 1980; Pintner & Paterson, 1917; Wallace & Corballis, 1973). It is important to note that this finding of a short span applies not only to the English materials (e.g., lists of words, letters, or digits), but also applies to studies that have measured serial recall of *signs*. Fairly typical results were found in the Bellugi et al. (1975) study, in which deaf adults' correct serial recall of signs reached an asymptote with a list length of four signs, while the hearing subjects reached asymptote with lists of six words. Thus, the differences in memory span found between hearing and deaf individuals appear to be due not simply to unfamiliarity with the English material; rather, they appear to be related to cognitive processes involved in short-term memory for linguistic materials, in general.

Ability to maintain a sequence of words in short-term memory is related to the use of phonological coding. That is, studies with orally trained subjects (Conrad, 1979), native signers of ASL (Hanson, 1982), and subjects mixed in terms of their educational and linguistic backgrounds (Lichtenstein, 1985), have all found strong correlations between the magnitude of the rhyme effect for deaf subjects and measured memory span. In these studies, the larger the rhyme effect for a deaf subject, the larger that subject's memory span. In contrast, no correlation between use of manual coding and measured memory span has been established (Lichtenstein, 1985).

Given this relationship between serial recall ability and phonological coding, we have suggested that one reason the skilled deaf reader uses phonological coding may have to do with the critical syntactic role played by sequential structuring in English (Hanson, 1982; Lake, 1980; Lichtenstein, 1985). This analysis suggests that an issue to be faced by teachers is how to educate deaf students to process a highly temporally structured language such as English.

Deaf readers and phonology

It is notable that the subjects in the studies discussed were not generally from oral backgrounds. In some cases, subjects were expressly selected because they were *native signers of ASL*. Yet, even these subjects, if skilled readers, were found to be using phonological information in the reading of English, rather than referring to ASL.

A discussion of phonological sensitivity in deaf readers always leads to the question of how this sensitivity is acquired. It is likely that congenitally, profoundly deaf readers acquire phonology from a combination of three sources: experience with the orthography through reading, experience in speaking, and experience in lipreading. In many of the studies discussed here, there was evidence of phonological processing for deaf subjects whose speech was not intelligible. That even these subjects use phonological coding suggests that deaf individuals' ability to use phonological information when reading is not well reflected in the intelligibility ratings of their speech. Further research is needed to determine the type of language instruction capable of promoting access to the speech skills most relevant to reading.

When this chapter was first planned, it was titled "Is reading different for deaf individuals?" The answer appears to be both yes and no. Clearly the answer is yes in the sense that deaf readers will bring to the task of reading very different sets of language experiences than the hearing child. These differences will require special instruction. But, the answer is also no. The evidence indicates that skilled deaf readers use their knowledge of the structure of English when reading. Although sign coding, in theory, might be used as an alternative to phonological coding for deaf signers, the research using various short-term memory and reading tasks has found little evidence that words are processed with reference to sign by the better deaf readers. Rather, the better deaf readers, like the better hearing readers, have learned to abstract phonological information from the orthography, despite congenital and profound hearing impairment.

The finding of phonological processing by deaf readers, particularly deaf readers skilled in ASL, makes a strong case for the importance of phonological sensitivity in the acquisition of skilled reading, whether the reader is hearing or deaf. For deaf readers, the acquisition and use of phonological information is extremely difficult. They would be expected to use alternatives such as visual (orthographic) or sign strategy, if such were effective. Yet, the evidence indicates that the successful deaf readers do not rely on these alternatives.

REFERENCES

Battison, R. (1978). *Lexical borrowing in American Sign Language*. Silver Spring, MD: Linstok Press.

Bellugi, U., Klima, E. S., & Siple, P. (1975). Remembering in signs. *Cognition*, 3, 93-125.

Belmont, J. M., & Karchmer, M. A. (1978). Deaf people's memory: There are problems testing special populations. In

M. Gruneberg & R. Sykes (Eds.), *Practical aspects of memory* (pp. 581-588). London: Academic Press.

Blair, F. X. (1957). A study of the visual memory of deaf and hearing children. *American Annals of the Deaf, 102*, 254-263.

Burnet, J. R. (1854). The necessity of methodical signs considered. *American Annals of the Deaf, 7*, 1-14.

Conrad, R. (1979). *The deaf schoolchild*. London: Harper Row.

Erickson, D., Mattingly, I. G., & Turvey, M. T. (1977). Phonetic activity in reading: An experiment with Kanji. *Language and Speech, 20*, 384-403.

Furth, H. G. (1966). A comparison of reading test norms of deaf and hearing children. *American Annals of the Deaf, 111*, 461-462.

Hanson, V. L. (1982). Short-term recall by deaf signers of American Sign Language: Implications for order recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 572-583.

Hanson, V. L. (1989). *On the use of grapheme-phoneme correspondences by deaf readers*. Manuscript.

Hanson, V. L. (in press). Recall of order information by deaf signers: Cues to memory span deficits. *Memory & Cognition*.

Hanson, V. L. & Feldman, L. B. (1989). Language specificity in lexical organization: Evidence from deaf signers' lexical organization of ASL and English. *Memory & Cognition, 17*, 292-301.

Hanson, V. L., & Fowler, C. A. (1987). Phonological coding in word reading: Evidence from hearing and deaf readers. *Memory & Cognition, 15*, 199-207.

Hanson, V. L., Liberman, I. Y, & Shankweiler, D. (1984). Linguistic coding by deaf children in relation to beginning reading success. *Journal of Experimental Child Psychology, 37*.

Hanson, V. L., & McGarr, N. S. (1989). Rhyme generation by deaf adults. *Journal of Speech and Hearing Research, 32*, 2-11.

Hanson, V. L., Shankweiler, D., & Fischer, F. W. (1983). Determinants of spelling ability in deaf and hearing adults: Access to linguistic structure. *Cognition, 14*, 323-44.

Karchmer, M. A., Milone, M. N., Jr., & Wolk, S. (1979). Educational significance of hearing loss at three levels of severity. *American Annals of the Deaf, 124*, 97-109.

Krakow, R. A., & Hanson, V. L. (1985). Deaf signers and serial recall in the visual modality: Memory for signs, fingerspelling, and print. *Memory & Cognition, 13*, 265-72.

Kyle, J. G. (1980). Sign coding in short term memory in the deaf. In B. Bergman & I. Ahlgren (Eds.), *Proceedings of the First International Symposium on Sign Language Research*. Stockholm: Swedish National Association of the Deaf.

Lake, D. (1980). Syntax and sequential memory in hearing impaired children. In H. N. Reynolds & C. M. Williams (Eds.), *Proceedings of the Gallaudet Conference on Reading in Relation to Deafness*. Washington, DC: Division of Research, Gallaudet College.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1-36.

Liberman, I. Y. (1971). Basic research in speech and lateralization of language: Some implications for reading disability. *Bulletin of the Orton Society, 21*, 71-87.

Liberman, I. Y. (1973). Segmentation of the spoken word and reading acquisition. *Bulletin of the Orton Society, 23*, 65-77.

Lichtenstein, E. H. (1985). Deaf working memory processes and English language skills. In D. S. Martin (Ed.), *Cognition, education, and deafness: Directions for research and instruction* (pp. 111-114). Washington, DC: Gallaudet College Press.

Macdonald, C. J. M. (1988). *Can oral-deaf readers use spelling-sound information to decode?* Manuscript.

Mann, V. A. (1985). A cross-linguistic perspective on the relation between temporary memory skills and early reading ability. *Remedial and Special Education, 6*, 37-42.

McDermott, M. J. (1984). *The role of linguistic processing in the silent reading act: Recoding strategies in good and poor deaf readers*. Doctoral dissertation, Brown University, Providence, RI.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1974). Functions of graphemic and phonemic codes in visual word-recognition. *Memory & Cognition, 2*, 309-321.

Pintner, R., & Paterson, D. G. (1917). A comparison of deaf and hearing children in visual memory for digits. *Journal of Experimental Psychology, 2*, 76-88.

Poizner, H., Bellugi, U., & Tweney, R. D. (1981). Processing of formational, semantic, and iconic information in American Sign Language. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1146-1159.

Reynolds, H. N. (1975). Development of reading ability in relation to deafness. *World Congress of the World Federation of the Deaf: Full Citizenship for all Deaf People* (pp. 273-284). Washington, DC: National Association for the Deaf

Shankweiler, D., Liberman, I. Y., Mark, L. S., Fowler, C. A., & Fischer, F. W. (1979). The speech code and learning to read. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 531-545.

Smith, C. R. (1975). Residual hearing and speech production in deaf children. *Journal of Speech and Hearing Research, 18*, 759-811.

Stokoe, W. C., Jr., Casterline, D., & Croneberg, C. (1965). *A dictionary of American Sign Language*. Washington, DC: Gallaudet College Press.

Treiman, R., & Hirsh-Pasek, K. (1983). Silent reading: Insights from congenitally deaf readers. *Cognitive Psychology, 15*, 39-65.

Trybus, R., & Karchmer, M. (1977). School achievement scores of hearing impaired children: National data on achievement status and growth patterns. *American Annals of the Deaf, 122*, 62-69.

Tzeng, O. J. L., Hung, D. L., & Wang, W. S-Y. (1977). Speech recoding in reading Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 621-630.

Wallace, G., & Corballis, M. C. (1973). Short-term memory and coding strategies in the deaf. *Journal of Experimental Psychology, 99*, 334-348.

FOOTNOTES

*In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 69-89). Ann Arbor: University of Michigan Press. (1989).

†Now at IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, New York

[1]The term *phonology* need not be limited to use with spoken languages. In the case of American Sign Language, for example, the term phonology has been used to describe the linguistic primitives related to the visible gestures by the hands, face, and body of the signer. In the present chapter, however, phonology will be restricted in reference only to features related to spoken languages.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 180-194

Syntactic Competence and Reading Ability in Children*

Shlomo Bentin,† Avital Deutsch,† and Isabelle Y. Liberman††

The effect of syntactic context on auditory word identification and on the ability to detect and correct syntactic errors in speech was examined in severely reading disabled children and in good and poor readers selected from the normal distribution of fourth graders. The poor readers were handicapped when correct reading required analysis of the sentence context. However, their phonological decoding ability was intact. Identification of words was less affected by syntactic context in the severely disabled readers than in either the good or poor readers. Moreover, the disabled readers were inferior to good readers in judging the syntactical integrity of spoken sentences and in their ability to correct the syntactically aberrant sentences. Poor readers were similar to good readers in the identification and judgment tasks, but inferior in the correction task. The results suggest that the severely disabled readers were inferior to both good and poor readers in syntactic awareness, and in ability to use syntactic rules, while poor readers were equal to good readers in syntactic awareness but were relatively impaired in using syntactic knowledge productively.

Fluent reading involves a complex interaction of several parallel processes that relate visual graphemic stimuli to specific entries in the lexicon and combine the semantic and syntactic information contained in those entries to apprehend the meaning of sentences. Some of these processes relate to the decoding of the phonological code from print while others relate to the assignment of meaning to the phonological units. Although the decoding of the phonological code can, in principle, be based solely on "bottom-up" application of grapheme-to-phoneme transformation rules, it is well documented that this process is supported by "top-down" streaming of lexical knowledge and contextual information. The common denominator of the "bottom-up" and "top-down" processes in reading is that both are components of the human linguistic endowment (see Perfetti, 1985; Rozin & Gleitman, 1977).

This study was supported by the Israel Foundations Trustees to Shlomo Bentin. Isabelle Liberman was supported in part by National Institute of Child Health, and Human Development Grant HD-01994 to Haskins Laboratories. The useful comments of Anne Fowler are much appreciated. We gratefully acknowledge the cooperation of the principals and teachers of the integrative schools "Luria" and "Stoane" in Jerusalem, and of the School for Remediation Teaching in Hertzelia, Israel.

The relative contribution of context dependent (top-down) processes to visual word recognition is determined by many factors among which reader competence is particularly important. Although some authors have argued that as fluency develops, the reader increasingly relies on contextual information during word recognition (Smith, 1971), more recent studies have disconfirmed this hypothesis. For example, semantic priming facilitates lexical decisions more in children than in adults and more in younger than in older children (Schvaneveldt, Ackerman, & Semlear, 1977; West & Stanovitch, 1978). Similarly, context effects are greater in poor readers than in good readers both in lexical decision (Schwantes, Boesl, & Ritz, 1980) and naming tasks (Perfetti, Goldman, & Hogaboam, 1979; Stanovitch, West, & Feeman, 1981). Within the same subject, larger context effects occur when bottom-up processes are inhibited by degrading or masking the target stimuli (e.g., Becker & Killion, 1977; Massaro, Jones, Lipscomb, & Scholz, 1978; Meyer, Schvaneveldt, & Ruddy, 1975). These results imply that the increased role of higher-level contextual processes in visual word recognition is caused by a need to compensate for deficiencies in lower level, decoding processes,

such as those that occur with poor readers or degraded stimuli (Perfetti & Roth, 1981; Stanovitch et al., 1981).

The observation that context effects are larger in poor than in good readers does not imply, however, that poor readers are better or more efficient users of contextual information than good readers. In fact, the opposite may be true. Both Perfetti et al. (1979) and Schwantes et al. (1980) reported that when subjects are required to use the context for predicting the target word before seeing it, skilled readers do better than poor readers. The magnitude of semantic priming effects in visual word recognition may therefore be a rather poor indicator of the real ability to use contextual information. A relatively unbiased method for comparing good and poor readers' awareness of contextual information would be to evaluate context effects on word recognition in situations that either eliminate the need for decoding of print, as in auditory presentation, or that force all readers, regardless of skill, to use contextual processes to the same extent.

As a language in which efficient use of contextual information is essential for fluent reading, Hebrew provides an excellent channel through which to examine syntactic effects. In Hebrew orthography, letters represent mostly consonants, while vowels are represented by diacritical marks placed below, within, or above the letters. The vowel marks are usually omitted in writing except in poetry, holy scripture, and children's literature. Because different words may be represented by the same consonants but different vowels, when these vowels are absent up to seven, eight, or more different words may be represented by the same string of letters. In addition to being semantically ambiguous, these Hebrew homographs are also phonemically equivocal because the (absent) vowels of the words that are responsible for the different meanings vary from word to word. Therefore, fluent reading of "unvoweled" Hebrew requires heavy reliance on contextual information.

Reading instruction in Hebrew starts, as a rule, with the "voweled" orthographical system in which the diacritical marks are presented with the consonant letters. The vowels are gradually omitted from school texts starting at the beginning of the third grade. During the third grade, the children begin to learn to read without vowels. By grade four, they are expected to be fluent readers of unvoweled texts. Informal discussions with teachers, however, revealed that the transition from reading voweled to reading unvoweled material is not equally easy for all children. According to teachers, some children are good readers as long as the diacritical marks are present but are slow in acquiring the skill of reading without the vowels. Because without vowels the context of the sentence is a primary source of phonological constraints on reading, we suspected that the children in this group, although knowing the grapheme-to-phoneme transformation rules, do not (or can not) use contextual information efficiently. Thus, in spite of being phonologically skilled, those children may be poor readers. Thus, Hebrew may be a convenient medium through which to test the hypothesis that at least some deficient readers are less able than good readers to use context.

We decided to manipulate syntactic rather than semantic context in the present research. The main reason for this decision is that syntax is probably a more basic linguistic ability than semantics (see Chomsky, 1969) and less affected by reading experience (Lasnik & Crain, 1985). In addition, syntactic violations are more clearly defined than are manipulations of semantic association strength.

The effect of prior syntactic structure on the processing of a visually presented target word has been investigated in a number of studies. Lexical decisions regarding target words are faster when they are preceded by syntactically appropriate primes than when preceded by syntactically inappropriate primes (Goodman, McClelland, & Gibbs, 1981; Lukatela, Kostić, Feldman, & Turvey, 1983). Lexical decision (West & Stanovitch, 1986; Wright & Garret, 1984) and naming (West & Stanovitch, 1986) are facilitated when targets are syntactically congruent with previously presented sentence fragments relative to when targets follow a syntactically neutral context.

Although the syntactic context effect on word recognition is reliable, information on the relation between this effect and reading ability is comparatively scarce and controversial (for a review see Vellutino, 1979). Several studies report that poor readers are inferior to normal readers in dealing with complex syntactical structures in speech (Brittain, 1970; Bryne, 1981; Cromer & Wiener, 1966; Goldman, 1976; Guthrie, 1973; Newcomer & Magee, 1977; Vogel, 1974). Other authors, however, have challenged a simple interpretation of these results (Glass & Perna, 1986). For example, Shankweiler and Crain (1986) have suggested that the poor readers' apparent deficiency in processing complex syntactic

information may be an epiphenomenon of limitations of the working memory processor, rooted in a difficulty in generating phonological codes.

The present study sought to examine the relation between word recognition and syntactic awareness; i.e., sensitivity to syntactic structure and the ability to use syntactic knowledge explicitly, in children who vary in reading competence. Experiment 1, using a group of adult fluent readers of Hebrew, was designed to establish the validity of the procedure of testing the effect of syntactic context on the identification of auditorily presented words masked by white noise. Experiment 2 examined syntactic context effects in two groups of children. One group was composed of children with learning disorders drawn from a population of students who were selected by the school system for special supplementary training in reading; a comparison group was formed of good readers drawn from the population of fourth-graders of two elementary schools. In Experiment 3, the same group of good readers was compared with a group of poor readers from the same elementary schools. The poor readers were matched with the good readers for their ability to apply grapheme-to-phoneme transformation rules in reading voweled pseudowords, but they were significantly inferior in reading sentences when the words were printed without the diacritical vowel marks.

EXPERIMENT 1

The purpose of the present experiment was to establish the effect of syntactic context on the identification of auditorily presented words that were masked by white noise. The auditory modality was used to attenuate the deficient readers' excessive reliance on contextual information in reading (which is presumably caused by the need to compensate for their difficulty in decoding the print). Stimulus masking was incorporated in the procedure because previous studies suggest that degradation increases the tendency of all subjects to use contextual information for word recognition (Becker & Killion, 1977; Stanovitch & West, 1981). We chose to use identification rather than reaction time measures in order to keep the measurement as simple and direct as possible.

Subjects were presented with a list of three- or four-word sentences that were pre-recorded on tape. In each sentence, white noise was superimposed on one or several (target) words. The subjects were instructed to identify the masked words. Half the targets in the list were congruent with the syntactic structure of the sentence in which they appeared whereas the other targets were incongruent, that is, caused a syntactic violation. We predicted that the percentage of correctly identified targets would be higher for syntactically congruent words than for syntactically incongruent words.

Method

Subjects

The subjects were 28 undergraduate students (14 males) who participated in the experiment for course credit or for payment. They were all native speakers of Hebrew with normal hearing.

Test Materials

The auditory test included 104 three- or four-word sentences. Each sentence was used in two forms: a) syntactically correct and b) syntactically incorrect. The incorrect sentences were constructed by changing the correct sentences in one of the following 10 ways.

Type 1 - In this category there were 12 sentences in which the gender compatibility between the subject and the predicate was altered. In six of these sentences a masculine subject was presented with a feminine predicate and in the other six a feminine subject was presented with a masculine predicate. The masked target was the predicate, which was always the last word in the sentence.

Type 2 - In this category there were 12 sentences in which the compatibility of number between subject and predicate was altered. In six of these sentences a singular predicate followed a subject in plural form, and vice-versa in the other six. The masked target was the predicate which was the last word in each sentence.

Type 3 - There were eight sentences in this category. The compatibility of gender and number between the subject and the predicate was altered in each sentence. The masked target was the predicate which was the last word in each sentence.

Type 4 - In Hebrew, prepositions related to personal pronouns (e.g., "on me") become one word. The violation in this category consisted of the decomposition of the composed pronoun into two separate words (the pronoun and the preposition) which kept the meaning but were syntactically incorrect. Eight different prepositions were used, four times each, totaling 32 sentences. The masked target words were the composed or decomposed pronouns. Because the

decomposition of the pronoun and preposition added one word to the sentence, another word was excluded from each of the incorrect sentences.

The syntactical violations of types 5 to 10 were based on changing compulsory order of parts of the sentence. Therefore, the sentences of these types were masked from the first to the last word.

Type 5 - In the ten sentences of this type, the order of the attribute and its nucleus was reversed.

Type 6 - In each of the six three-word sentences of this type, the predicate was incorrectly introduced between the subject and its attribute.

Type 7 - In Hebrew, the negation always comes before the negated predicate. We altered this fixed order in six sentences, using three different negation words.

Type 8 - In six sentences, the interrogative word was moved from its fixed place at the beginning of the sentence to the second place.

Type 9 - In six sentences, the fixed order of preposition and noun was reversed so that the noun appeared before the preposition.

Type 10 - In six sentences, the copula that should occur between the subject and the predicate was moved to the beginning or the end of the sentence.

All 104 correct and 104 incorrect sentences were recorded on tape by a female native speaker of Hebrew. The tapes were sampled at 20KHz. The masked intervals were marked and white noise was digitally added to the marked epochs with a signal-to-noise ratio of 1:2.75. This ratio was determined on the basis of pilot tests so that correct target identification level was about 50%.

The 208 sentences were organized into two lists. In each list, 52 sentences were syntactically correct and 52 sentences were syntactically incorrect. Each sentence appeared in each list only once, either in correct or incorrect form. Sentences that were correct in list A were incorrect in list B and vice versa. Fourteen subjects were tested with list A and the other 14 with list B. Thus, each subject listened to an equal number of correct and incorrect sentences, and across subjects each sentence appeared an equal number of times in each form.

The sentences in each list were randomized and re-recorded on tape. Subjects listened to the tapes via Semmheiser earphones (HD-420).

Procedure

The subjects were tested individually in a quiet room. The experimenter listened to the stimuli simultaneously with the subject and stopped the tape-recorder at the end of each sentence. The subjects were asked to repeat the masked part of each sentence, and were encouraged to guess whenever necessary. The subjects' responses were recorded manually by the experimenter. Subjects were randomly assigned to List A or B.

Results and Discussion

The average percentage of correct responses, across subjects and sentence types was 41.3%. Overall, correct identification was 67.0% for the syntactically correct sentences but only 15.6% for the syntactically incorrect sentences. The syntactic context effect was evident for each type of syntactic violation (Table 1).

Table 1. *Percentage of correct identification of syntactically correct and syntactically incorrect sentences in each violation-type category (see text).*

	Type of syntactic violation									
	1	2	3	4	5	6	7	8	9	10
Syntactically correct										
Mean	60.7	72.6	67.9	83.2	54.3	67.9	71.5	69.0	64.3	58.4
(SEm)	6.2	4.4	4.5	1.7	4.6	6.6	7.1	6.9	6.1	7.0
Syntactically incorrect										
Mean	18.7	25.6	2.7	20.6	12.2	10.7	14.3	8.2	19.0	23.7
(SEm)	4.9	5.0	2.6	5.2	4.2	6.9	6.6	3.4	7.6	6.5

These observations were confirmed by two-factor analyses of variance with subjects and sentences as random variables. The factors were syntactic context (correct, incorrect) and syntactic violation type (Type 1 to 10). The main effect of syntactic context was significant ($F(1,26)=588.44$, MSe=629, $p < .0001$ for the subject-analysis and $F(1,91)=140.93$, MSe=624, $p < .0001$ for the stimulus-analysis). The main effect of sentence type was also significant ($F(9,234)=4.5$, MSe=404, $p < .0001$ for the subject-analysis and $F(9,91)=2.21$, MSe=437, $p < .03$ for the stimulus-analysis). The context effect was conspicuous for all syntactic violation types, but, as suggested by an interaction between the two factors, its magnitude differed. This interaction was significant for the subject analysis ($F(9,234)=4.73$, MSe=322, $p < .0001$), but only marginal for the stimulus analysis ($F(9,91)=1.86$, MSe=624, $p < .07$). Tukey-A post hoc analysis of the interaction revealed that the context effect was greater for type 3 (gender and number), type 4 (composite pronoun), type 8 (translocation of interrogative word), type 6 (separation of subject and attribute), and type 7 (translocation of negation word) than for all the other sentence-types. Within these two groups, the context effects were similar.

The results of Experiment 1 demonstrate that identification of words in sentences is influenced by the syntactic coherence of the sentence. Because, across subjects, exactly the same words were masked and had to be identified in the syntactically correct and incorrect sentences, the difference in the correct identification rate between the two modes of presentation is probably due to the manipulation of syntactic coherence. The magnitude of this effect seemed to vary across different types of syntactic anomalies but it was reliable and statistically significant for each type. Because we had neither a priori predictions about the effects of particular violation-types on identification nor clear post hoc explanations for the observed differences, and because the type of violation is not directly relevant to the issues investigated in this study, the syntactic violation-types will be collapsed in all further analyses.

On the basis of these results, we can use the technique of Experiment 1 to assess possible differences in the magnitude of the effect in good and poor readers.

EXPERIMENT 2

Experiment 2 compared the magnitude of the syntactic context effect on the identification of auditorily presented words in children who are good readers and children with a severe reading disability. As was elaborated in the introduction, although several studies reported that poor readers are deficient in syntactic comprehension (see Vellutino, 1979), others could not find solid evidence to support this hypothesis (e.g., Glass & Perna, 1986). If disabled readers are less aware of the syntactic structure of the sentence (as part of their general linguistic handicap), or do not use syntactic information as efficiently as good readers, syntactic context effects should be weaker in disabled than in good readers. Consequently, the effect of syntactic congruity on correct identification of sentences should be smaller in disabled than in good readers.

A second prediction concerns the nature of errors made by good and disabled readers in the identification of words presented in syntactically incorrect sentences. The auditory mask probably induces some degree of uncertainty in the auditory input. If listeners are aware of the sentence context, they may attempt to use it to complement the information that is missing in the auditory stream. Such a strategy would cause errors in the identification of words that violate the syntactical structure because in those sentences the target does not conform to the expected syntactic rules. Therefore, errors in identification that are induced by syntactic awareness should reflect the use of correct syntactic forms. In an English example, if the sentence were "I would like to have many child" and the word "child" was masked, the subject may erroneously identify the target as "children." On the other hand, if the subjects are not aware of the syntactic structure or not bothered by its violations, their errors in the identification of masked words should not be related to the syntactically correct form of the target. In this case, the response may be a randomly selected word, or may relate to the acoustical form — for example, substituting "mild" for the target "child" in the above example. If good readers are more aware of the syntactic structure of the sentence than disabled readers, the percentage of errors of the first type—"syntactic corrections," and of the second type—"random errors"—should vary with reading ability. In an extreme case, we should find more "syntactic correction" errors than "random" in good readers and vice versa in disabled readers.

Analyses of the errors made by each reading group would be a first step towards understanding the cause of inter-group differences in syntactic context effects, if they exist. However, in order to assess syntactic awareness as a metalinguistic

ability rather than automatic use of syntactic structures for word identification, a more direct measure had to be employed. In a recent study, Fowler (1988) compared good and poor readers' ability to detect and to correct violations of syntax in orally presented sentences. In that study, the ability to judge sentences as correct or incorrect in the "judgment" task was not associated with reading ability. In contrast, good readers performed syntactically better than poor readers in the "correction" task. Fowler concluded that poor readers do not differ from good readers in syntactic knowledge but that they may be inferior in manipulating verbal material in short-term memory (see also Shankweiler & Crain, 1986). We used Fowler's technique to supplement our study of syntactic context effects on the identification of orally presented words. If, as in Fowler's study, a difference emerges only for the correction condition, then syntactic awareness is not at fault. Rather, one would ascribe the differences to syntactic processing difficulties that prevent the disabled readers from using their syntactic knowledge productively.

Method

Tests and Materials

A. Reading Tests. We were interested in testing two kinds of reading: the ability to decode the phonology from print and the ability to use the sentence context in reading without vowels. Because all the standard reading tests in Hebrew primarily test reading comprehension, we constructed two new reading tests for our purposes. The first was a test of decoding ability. It contained a set of 24 meaningless three- or four-letter strings (pseudowords) presented with vowel marks. The vowels were chosen according to Hebrew morphophonemic rules, and included all lawful combinations. Each pseudoword was printed individually on a white, 9 cm X 12 cm cardboard. The size of each letter was 0.5 cm. The subject was instructed to read each pseudoword exactly as it was written. The accuracy and naming onset time were measured. The subject's score on this test consisted of the percentage of accurately read pseudowords and the mean latency of naming onset time.

The second test was designed to test the ability to read Hebrew without vowel marks and particularly to use the sentence context to determine the reading of unvoweled Hebrew words that were both phonologically and semantically ambiguous. This test contained 48 four- or five-word sentences printed on white cardbord using the same fonts as for the pseudowords in the former test. The last word in each sentence was the target word. In the absence of vowel marks, 32 out of the 48 targets were phonologically ambiguous, i.e., they could have been assigned at least two sets of vowels to form two different words. Thus, correct reading of those targets could be determined only by apprehending the meaning of the sentence. The 32 ambiguous targets were 16 pairs of identical letter strings each representing a different word in the respective sentence. Eight of these 16 ambiguous targets represented two words of equal frequency. The words represented by each of the remaining ambiguous words differed in frequency such that one member of each pair was a high-frequency word while the other member was a low-frequency word. In a previous study Bentin and Frost (1987) reported that when undergraduates were presented with isolated ambiguous words in a naming task, they tended to choose the most frequent phonological alternative. We assumed that, without context, the children would tend to choose the same. Therefore, insensitivity to the context of the sentence should increase the number of errors in reading the targets, particularly when the correct response requires the use of the less frequent phonological alternative. The remaining 16 targets were words that without the vowel marks could have been meaningfully read in only one manner. Eight of those sixteen targets were high-frequency words and the other eight targets were low-frequency words. The subjects were instructed to read each sentence aloud. The time that elapsed from the moment the sentence was exposed until the subject finished reading it was measured to the nearest millisecond. The score on this test was the average percentage of errors and the average time to read a sentence.

In addition to these two special purpose tests, each subject was tested for reading comprehension by a standard test. The nation-wide average score on this test for fourth graders is 70% with a SD of 12%.

B. Intelligence tests. The IQ of each subject was obtained either using the WISC (Full Scale) (whenever those data were available) or testing the children on the Raven Colored Matrices and transforming their performance into IQ scores.

C. Syntactic awareness test. Syntactic awareness was assessed by testing identification of auditorily presented words masked by white noise as in Experiment 1. On the basis of a pilot

study with children, in order to keep the overall correct identification of targets around 50%, we increased the signal-to-noise ratio in Experiment 2 from 1:2.75 to 1:2.25.

Procedure

Each child was tested individually in three sessions. During the first session, reading performance and intelligence were tested. Reading performance was recorded on tape for subsequent error analysis and off-line measuring of time. At the end of the test of reading without vowels, the experimenter verified whether the subject knew the meaning of the targets that had been read incorrectly. In the very few doubtful cases, the sentence was excluded and a substitute sentence of the same type was given. The children who had been selected for this study (see below) were invited to a second session during which the auditory word identification test was given. The procedures for the word identification test were identical to those of Experiment 1. In addition, at the end of the second session, the children were tested for the ability to repeat from memory the sentences presented during the auditory test. This was done by presenting the children with 16 sentences selected from the same pool of sentences from which the test set was selected. Eight of these 16 sentences were syntactically correct and the other eight syntactically incorrect. In the repetition test the sentences were presented without the masking noise. Finally, during a third session (three months later), all 104 sentences were presented to each child without the masking noise. Following the presentation of each sentence the child was asked whether "this is the way it should be said in Hebrew" (Judgment Task). Whenever the answer was "no," the child was asked to correct the sentence (Correction Task).

Subjects

The good readers were 15 children (7 males) selected from a population of fourth graders of two elementary schools in Jerusalem. Their ages ranged between 8.9 and 9.7 years (mean age 9.3 years). The average IQ (FS) score (as assessed by transforming the Raven score) was 102.5, ranging from 85 to 122.5. They were selected to match poor readers from the same school on decoding ability and IQ. The precise selection criteria will be elaborated in Experiment 3.

The disabled readers were 19 children (12 males), aged from 9.7 to 14 years, (mean age 11.6), selected from a population of 32 children with severe reading disorders who had been referred for special supplementary training in reading. They were within the normal intelligence range (Mean IQ (FS)=104.83, ranging between 85 and 130). The disabled readers selected for the present study were chosen because they not only showed poor decoding ability, as compared to good readers, but also performed badly in the test of reading without vowels, thus suggesting special problems in dealing with context. Table 2 presents the reading performance of the good and deficient readers as revealed by our reading tests.

Table 2. *Reading performance of the severely disabled and good readers.*

	Reading voweled Nonwords		Reading unvoweled Sentences		Reading Comprehension
	% of errors	Time per item	% of errors	Time per sentence	% correct
Good Readers	8.1%	1.6 sec	4.1%	2.9 sec	81.3%
Disabled Readers	38.8%	2.6 sec	19.6%	6.3 sec	55.7%

Children in both groups were all native speakers of Hebrew without known motor, sensory, or emotional disorders. All children had been tested for normal hearing.

Results

The overall percentage of correct identification of masked targets was similar in good (44.2%) and disabled readers (48.3%) ($F(1,32)=2.52$, MSe=112, $p > .12$). However, the percentages of syntactically correct and incorrect sentences that were correctly identified were different in the two groups (Table 3). Although syntactically correct sentences were identified better than syntactically incorrect sentences in both groups, the effect of the syntactic context was smaller in the disabled readers than in the good readers.

This observation was supported by a mixed-model two-factors analysis of variance. The between-subjects factor was reading ability, and the within-subjects factor was syntactic context (correct, incorrect). The syntactic context effect was highly significant across groups ($F(1,32)=784.47$, MSe=50, $p < .0001$). A more interesting result, however, was the significant interaction that revealed that the syntactic context effect was greater in good readers than in disabled readers ($F(1,32)=11.90$, MSe=50, $p < .002$).

Post hoc analysis revealed that good and disabled readers performed equally well with correct sentences. However, the percentage of correct identifications of words embedded in incorrect sentences was higher in disabled than in good readers ($p < .01$).

Table 3. *Percentage of correct identification of syntactically correct and incorrect sentences in the reading disabled and good readers.*

	Good Readers	Disabled Readers
Syntactically Correct		
Mean	71.4	69.5
(SEm)	2.2	1.9
Syntactically Incorrect		
Mean	17.0	27.0
(SEm)	2.4	2.2

The errors that children made were distributed among four error types: Type 1 errors were "syntactical corrections," that is, errors that were made in attempt to use the correct syntactic structure of a syntactically incorrect sentence. Type 2 errors were "random errors"—misidentifications that made no sense whatsoever or reflected acoustical confusions. Type 3 errors were "logical substitutions," that is, substitutions of the masked words with other words that gave the sentence a logical meaning. Type 4 were "I don't know" responses, which were not encouraged but were accepted. The percentage of errors of each type (out of the total number of responses) in each group is presented in Table 4.

Table 4. *Percentage of errors of each type (out of the total number of responses) made by the disabled and good readers in the auditory identification task.*

	Type of error			
	Corrections	Random	Logical Substitutions	"I don't know"
Reading Disabled	3.9	10.2	21.9	15.7
(SEm)	0.4	1.5	1.7	2.2
Good Readers	6.2	3.3	17.4	28.8
(SEm)	0.8	0.7	1.7	2.8

Because we had clear predictions only for syntactic corrections and random errors, we analyzed the distribution of these two error types in each group by a mixed-model (reading group X error type) analysis of variance. This analysis showed that, across the two types of error, the good readers made fewer errors than the disabled readers ($F(1,32)=5.11$, $MSe=18$, $p < .031$). Across groups, the percentage of errors of each type was similar ($F(1,32)=3.01$, $MSe=16$, $p > .09$). Most interesting, the interaction between reading ability and error type was highly significant ($F(1,32)=21.54$, $MSe=16$, $p < .0001$). Post hoc analysis (Tukey-A) revealed that more random errors were made by disabled readers than by good readers, whereas syntactic corrections were more frequent in good than in disabled readers. All the children were able to repeat verbatim all sixteen sentences that they heard without the masking noise.

Good readers were better than disabled readers on both the judgment and the correction tests. Among the disabled readers, however, a secondary distinction was evident between four children who were 13-14 years old and those who were younger. The mean percentage of errors made by each group in each task is presented in Table 5.

Table 5. *Percentage of errors made by disabled and good readers in the judgment and correction tasks.*

	Task	
	"Judgment"	"Correction"
Good Readers	1.3	5.4
(SEm)	0.4	0.9
Reading Disabled	6.9	32.1
(SEm)	1.9	4.5
Older reading Disabled	0.3	5.6

Because the number of older disabled readers was too small to form a reliable independent level in a factorial design but, on the other hand, clearly formed a distinct group, they were excluded from the statistical evaluation. Thus, the percentage of errors in each task was compared only for good and disabled readers who were more similar in chronological age. As before, a mixed-model analysis of variance was employed where the between-subjects factor was reading group and the within-subject factor was the test

(judgment or correction). The analysis of variance showed that good readers made significantly fewer errors than disabled readers ($F(1,24)=26.58$, MSe=127, $p < .0001$) and that more errors were made in the correction than in the judgment task ($F(1,24)=79.25$, MSe=35, $p < .0001$). A significant interaction suggested that the task affected the percentage of errors made by disabled readers more than it affected the good readers ($F(1,24)=41.0$, MSe=35, $p < .0001$). This interaction supports Fowler's results by emphasizing the difference between the judgment and correction tests. However, in contrast to her results, post hoc Tukey-A tests revealed that the good readers made significantly fewer errors than disabled readers not only in the correction task, but also in the judgment task. The inclusion of the four older disabled readers in the analysis did not change the pattern of results, although these four children clearly performed better than the other disabled readers.

Discussion

For children with a severe reading disability, the syntactic context effect on the identification of spoken words was smaller than for good readers. One explanation for the results might be that disabled readers are worse at identifying auditorily masked words than good readers (Brady, Shankweiler, & Mann, 1983). Such an hypothesis, however, is not supported by the present data. If the disabled readers in the present study had been handicapped in the identification of masked words, any manipulation that increased the difficulty of identifying the words should have had a greater effect on disabled than on good readers. Therefore, we should have observed a stronger rather than a weaker syntactic context effect in disabled readers. Further, if masking had a more deleterious effect on identification of words by disabled relative to good readers, the overall identification performance in the disabled group should have been lower. In fact, the overall correct identification percentage in disabled readers was slightly higher than in the good reader group.

A second account of the results might be that the smaller syntactic context effect in poor readers reflects a more general problem, such as disorders of short-term or working memory. There is indeed ample evidence that disabled readers have problems with verbal short-term memory (Mann, Liberman, & Shankweiler, 1980; for a review see Brady, 1986). Therefore, memory disorders might explain why their performance is affected by sentence context less than that of good readers even when decoding difficulties are eliminated; they simply do not remember the sentence well enough. However, a simply reduced short-term memory span cannot easily account for the present results because the children in both reading groups could accurately repeat sentences similar to those used in the identification task without any difficulty. It is still possible, however, that more complex working memory problems could have contributed to the disabled readers' pattern of performance. We will return to this hypothesis in the General Discussion.

We are left with the most direct hypothesis that inferior syntactic awareness is the reason for the relatively poor use of syntactic context by the reading disabled children of Experiment 2. This possibility is supported by the results of the error analysis. The percentage of "syntactic correction" errors made by good readers was almost twice as great as that made by disabled readers. "Syntactic correction" errors could have only been made when the subject knew what the correct structure of the sentence should have been and expected it. In those circumstances, when the physical stimulus was degraded the good readers applied syntactic rules and misidentified the target. In the same situation, getting only partial information from degraded stimuli, disabled readers often applied a random guessing strategy disregarding the sentence context completely. Indeed, the percentage of "random" errors was three times greater in the disabled readers group than in good readers.

An additional question examined in the present experiment was whether the disabled readers had mastered the correct syntactic structures but did not use them properly, or had problems with basic syntactic knowledge. We examined this question by testing the ability of both groups to detect violations of syntactic structure and to correct the detected violations. The good readers performed better than the disabled readers in both tasks. Although the difference between the groups was greater for the correction task, disabled readers were significantly inferior to good readers in the judgment task as well. This latter result contradicts the results reported by Fowler (1988) and suggests that this group of disabled readers were inferior to good readers in their awareness of basic syntactic structures.

The discrepancy between the present results and Fowler's (1988) results, as well as the disagreement between our conclusion regarding the syntactic awareness of disabled readers and

previous assertions in the literature that basic phonological disability and deficient use of working memory mechanisms underlies the syntactic inferiority observed in poor readers (e.g., Shankweiler & Crain, 1986; Shankweiler et al., in press), can be explained in two ways. One possible explanation is that different types of mechanisms underly reading deficiencies in different languages. Recall that we selected our deficient reader group to emphasize problems of using context while reading without vowel marks. In doing so, we may have selected a group of children who were poor in syntactic processing. A second explanation is that we have examined children with a reading disability that was considerably more severe than that of the poor readers examined by Fowler. Experiment 3 therefore attempted to generalize the results of the present experiment to poor readers selected, as in Fowler's study, from the normal student population of regular elementary schools.

EXPERIMENT 3

Any attempt to generalize about the characteristics of reading disability or to predict the performance of children with reading disorders is impeded by the heterogeneity of this population. Indeed, reading disorders can appear as the most conspicuous symptom in children who suffer from attentional disorders or general learning disability; they can be the main symptom (but rarely the only symptom) of developmental dyslexia and, at the other extreme, they may characterize the performance of otherwise normal students who happen to be at the lower end of a normal distribution of reading ability. It is possible, therefore, that the prior selection of different types of reading disorders underlies most disagreements about this important handicap among educators and scientific investigators.

In Experiment 2, the reading disabled children were selected from a population of children with severe reading disorders. Although they were at least in the fourth grade, had normal IQ's, and had no documented neurological symptoms, some of those children could hardly read single words with or without vowel marks. We found that they were inferior to good readers in syntactic knowledge and in using syntactic context to help identify spoken words. In Experiment 3, our aim was to extend these findings to another reader group. Thus, we compared good readers with children in regular classes who, when formally tested, were inferior to good readers in reading performance. In particular, we wished to compare the good readers with a group of relatively poor readers who were equal to the good readers in basic decoding ability (as revealed by their performance on reading voweled pseudowords) but were poorer at reading without vowel marks. We assumed that the relatively poor reading performance of this group primarily reflects inefficient use of the sentence context, and expected to be able to measure this relative disability by our auditory test of syntactic context effects. In addition, the good and poor readers in the present experiment were tested for ability to detect and to correct syntactic violations.

Method

Subjects

The subjects were 30 children selected from a population of 167 fourth graders in two public elementary schools. The selection was based on performance on the test of decoding ability and the test of reading without vowels which were described in Experiment 2. Two reading groups were assembled. The poor reader group included 15 children (9 males); their ages ranged between 8.8 and 9.6 years (mean age 9.1 years). Their average IQ (FS) score (as assessed by transforming the Raven score) was 102.5, ranging from 85 to 122.5. Each of the those poor readers made no more than four errors (16.6%) in the test of decoding voweled pseudowords but at least twice as many errors as the good readers while reading meaningful sentences without vowels. On the basis of the assumption that the relatively poor reading performance of those children reflected problems with processing of contextual information, we will label this group the "Poor context" group. The good readers were the same 15 children who were described in Experiment 2. Each child in this group was selected to match one child in the poor context group on the ability to decode voweled pseudowords and on IQ. However, the good readers performance on the unvoweled sentences was at least twice as good as that of his or her matched subject in the poor context group. The average scores of the two reading groups on the reading tests are presented in Table 6.

Tests and Materials

The reading tests and the auditory word identification test were identical to those described in Experiment 2. The IQ scores were estimated by testing the children with the Raven Colored Matrices test.

Table 6. *Reading performance of the good and poor context readers.*

	Reading voweled Nonwords		Reading unvoweled Sentences		Reading Comprehension
	% of errors	Time per item	% of errors	Time per sentence	% correct
Good Readers	8.1%	1.6 sec	4.1%	2.9 sec	81.3%
Poor Context Readers	7.8%	1.7 sec	18.2%	4.7 sec	72.0%,

Procedure

The procedure was similar to that employed in Experiment 2. The children were tested in two sessions. The first session was dedicated to the selection of subjects for this study. All 167 fourth-graders were tested for reading ability and IQ. During the second session only the selected children were tested on the auditory word identification test, and during a third session, their ability to detect and correct the syntactic violations. During the third session the sentences were presented without any masking noise. Sessions one and two were held close to the beginning of the academic year. Session three was three months latter.

Results

The percentage of total correct identifications in the "poor context" group (40.4%) was not significantly different from that of the good readers (44.2%) ($F(1,28)=1.03$, $MSe=209$, $p > .31$).

The percentages of syntactically correct and incorrect sentences that were correctly identified in each group are presented in Table 7.

Table 7. *Percentage of correct identification of syntactically correct and incorrect sentences in good and poor context readers.*

	Good Readers	Poor Context Readers
Syntactically Correct		
Mean	71.4	65.3
(SEm)	2.2	3.9
Syntactically Incorrect		
Mean	17.0	15.4
(SEm)	2.4	3.6

These data were analyzed by a mixed-model analysis of variance as in Experiment 2. As before, the syntactic context was highly significant ($F(1,28)=505.56$, $MSe=81$, $p < .0001$). However, in contrast to the findings of Experiment 2, the interaction between the syntactic context effect and reading group was not significant ($F(1,28)=0.96$).

As in Experiment 2, the errors made by each group were categorized into four types. Type 1 were "syntactical corrections," Type 2 were "random errors," Type 3 were "logical substitutions," and Type 4 were "I don't know." The distribution of errors in each of the two reading groups (out of the total number of responses) is presented in Table 8.

Table 8. *Percentage of errors of each type made by good and poor context readers (out of the total number of responses) in the auditory identification task.*

	Type of error			
	Corrections	Random	Logical Substitutions	"I don't know"
Good Readers				
Mean	6.2	3.3	17.4	28.8
(SEm)	0.8	0.7	1.7	2.8
Poor Context Readers				
Mean	3.5	6.9	24.5	24.8
(SEm)	0.6	1.2	2.9	3.5

Our a priori predictions concerned only errors of Type 1 (correction) and Type 2 (random errors). A mixed-model analysis of variance showed no significant main effects but a significant interaction between the type of error and reading group ($F(1,28)=10.63$, $MSe=14$, $p < .003$). Post hoc comparisons (Tukey-A) showed that the percentage of syntactical correction errors was higher in good readers than in the poor context group, whereas the percentage of random errors was higher in the poor context group than in good readers.

The average percentages of errors in the judgment and correction tasks for each group are presented in Table 9.

A mixed-model analysis of variance as in Experiment 2 was used to analyze these data. Across groups, there were more errors in the correction task than in the judgment test ($F(1,24)=39.16$, $MSe=16$, $p < .0001$). Overall, the good readers made fewer errors than poor context readers ($F(1,24)=5.24$, $MSe=25$, $p < .035$). The

interaction between the test and the group factors was significant ($F(1,24)=6.32$, MSe=16, $p < .020$). Replicating Fowler's (1988) results, post hoc analysis revealed that good and poor context readers did not differ in the judgment test, whereas in the correction test good readers made significantly fewer errors than poor context readers.

Table 9. *Percentage of errors made by good and poor context readers in the judgment and correction tasks.*

	Task	
	"Judgment"	"Correction"
Good Readers		
Mean	1.3	5.4
(SEm)	0.4	0.9
Poor Context Readers		
Mean	1.7	11.4
(SEm)	0.5	2.3

All children were able to repeat verbatim all the 16 sentences presented to them in absence of masking white noise.

Discussion

Experiment 3 sought to generalize the results of Experiment 2 to groups of relatively poor readers selected from the normal distribution of fourth graders. Unlike in Experiment 2, the magnitude of the syntactic context effect was similar in the good and in the poor context readers.

The syntactic ability of the children in the poor context group was not, however, entirely equivalent to that of the good readers. Analysis of the identification errors made by each group revealed that the proportion of errors that reflected an attempt to correct the syntactic violation was lower in children who had relatively more difficulties in reading unvoweled words than in good readers who were matched with them for phonological decoding ability. In contrast, the proportion of misidentifications that reflected total ignorance of the sentence's context (either syntactic or semantic) was lower in good readers than in the poor context group. This pattern of errors might suggest that although word identification was similarly affected by syntactic context in both reading groups, the good readers were more aware of the syntactic structure of the sentence than were the children in the poor context group. The results of the judgment test, however, did not support this hypothesis. As it turned out, both groups were equally sensitive to violations of syntactic structures. It is possible though, that part of this result reflected a ceiling effect in that task. The groups differed, however, in their ability to correct those violations.

The common aspect of both the "syntactical correction" errors and the test of correcting syntactic violations is that both measures reflect the child's ability to actively generate correct syntactic structures. This ability is not required by the judgment test and may not be reflected in identification performance. Therefore, the present data suggest that although the good and the relatively poorer readers did not differ in their syntactic awareness—that is, in the sensitivity to and knowledge of basic syntactic structures—the good readers had a superior ability to use their syntactic knowledge, and a tendency to do so.

GENERAL DISCUSSION

In the present study we examined the relation between reading ability and syntactic competence as it is reflected in the ability to use syntactic context for word identification and to detect and correct syntactic violations. In contrast to the great majority of studies of context effects in good and poor readers, we used auditory rather than printed word identification. Auditory presentation was used to circumvent a bias that might have been induced by the reading disorder itself. Thus, we were better able to assess differences in syntactic processing ability that might relate to reading achievement. The sensitivity of our auditory test to syntactic context was verified by showing that undergraduate students, fluent readers of Hebrew, identified target words masked by white noise significantly more accurately if the targets were syntactically congruent with the sentence in which they appeared than if they violated the syntactic structure.

A syntactic effect similar to that found in undergraduates was obtained when the same test was given to fourth graders. However, the difference between the correct identification of syntactically correct and syntactically incorrect sentences was smaller in a group of children with a severe reading disability than in either good readers or relatively poor readers selected from the normal distribution of fourth-grade students (the poor context group). The good readers and the poor context group did not differ in the auditory identification test.

A second difference between the severely reading disabled and the children in the poor

context group was observed in the judgment task. Children in the poor context group detected sentences that contained an error as well as good readers. In contrast, the reading disabled were worse in this test than either the good readers or the poor context readers.

The relative inferiority of the severely disabled readers can not be accounted for only by a simple reduction of their short-term memory span. In contrast to the complex sentences and complex syntactic structures typically used in other studies, we used only very short and simple sentences (three or four words). When formally tested, all the children were able to repeat the sentences verbatim without any problem. Holding a sentence in working memory for syntactic analysis probably requires more mental effort and retention of the whole sentence for a longer time than required by immediate repetition. As was previously reported, the factor of delay influences the memory ability of poor readers more than that of good readers (Liberman, Shankweiler, Liberman, Fowler, & Fisher, 1977). However, rather than requiring the manipulation of more subtle syntactic aspects, the syntactic violations which we have used in the present study were, as we have said, straightforward corruptions of the basic syntactic relationship between subject and predicate or a word order that clearly violated the syntactic structure of the sentence. Therefore, we agree with Byrne (1981) in doubting that deficient use of verbal memory mechanisms by disabled readers, at least as this deficiency could be revealed by simple repetition, was a major cause for the deficient use of syntactic context in the present study. Instead, we are inclined to believe that the reduced syntactic ability suggested by the performance of disabled readers reflected a genuine deficiency of linguistic endowment (in the syntactic and phonological domains) rather than reduced general cognitive ability or poor metalinguistic insight.

Although the syntactic context effect on the identification task was equal in the good readers and the poor context readers, the syntactic competence of these two groups was not entirely equivalent. In particular, the good readers made significantly more syntactical correction errors than the poor context readers. The difference between the two groups was even more conspicuous in the correction task. Similar to the results reported by Fowler (1988) for American poor readers, the ability of poor context children to correct syntactic violations was significantly inferior to that of good readers. This result is particularly interesting because, as was noted earlier, the syntactic violations used in the present study were much simpler and more direct than those used by Fowler. Moreover, our samples of good and poor context readers were matched for their ability to decode and read voweled nonwords. Therefore, just as for the disabled readers, the difference between the ability of good and poor context readers to correct syntactically incorrect sentences cannot be easily accounted for only by assuming differences between the poor and the good readers in general cognitive skills. Rather, we suggest that, at least for the specifically selected group of poor readers whose reading errors reflected reduced ability to analyze contextual information, both the correction test and the pattern of errors in the identification test suggest a specific impairment in the ability to use their syntactic knowledge in a productive way.

Although both the reading disabled and the poor context readers are inferior to good readers in syntactic competence, these two groups differ from one another. In comparison to good readers, the disabled readers showed a weaker syntactic context effect in the word identification task, an inferior ability to detect syntactical aberrations in spoken sentences, and an inferior ability to correct detected syntactically incorrect sentences. The poor context children were equal to good readers in the syntactic context effect on word identification, were equally able to detect syntactic aberrations, but were inferior to good readers in the ability to correct the detected errors. This pattern of results suggest that the different tasks tap different aspects of syntactic competence which might develop at different rates.

Some insight into the nature of the syntactic disability reflected by the word identification task comes from the observation that the significant interaction between the reading group and the syntactic context effect was not caused by a symmetrical effect of reading group on both syntactically correct and incorrect sentences. Rather, it seems that syntactically correct sentences were identified equally well by both reading disabled and good readers; however, good readers were affected more than reading disabled by violations of the syntactic structure of sentences. A possible interpretation that is supported by these data is that automatic syntactic processing was equivalent in both groups, but that the disabled readers were less aware of the syntactic structure and did not use identification strategies that were based upon it.

A more definite interpretation obviously requires a neutral condition in the identification task, which was absent in this study. However, these results strongly suggest that the identification test is sensitive more to strategic differences and syntactic awareness than to (automatic) syntactic processing.

It is noteworthy to recall in this connection the four older disabled readers: Although they were similar to other disabled readers in the auditory identification task, their performance in the judgment and correction tasks was similar to that of good readers. These older children may have had a higher level of syntactic competence so that when their attention was intentionally directed to the structure of the sentence (as was the case in the judgment and correction tasks in contrast to the identification task), the additional knowledge enabled them to use their syntactic knowledge productively.

In conclusion, the results of the present study suggest that syntactic factors are directly related to reading disabilities, at least in Hebrew. Two distinct populations of poor readers have been identified. One group was formed of children who in absence of a better term were labeled reading disabled. These children were probably able to use basic syntactic structures, as was evident in their everyday speech ability and in their identification of syntactically correct sentences. However, they were not explicitly aware of the syntactic structures, and therefore were not inhibited by semantic incongruity in the identification test; they were less able than good readers to detect syntactically incorrect sentences, and they were less able to correct those errors that had been detected. The second group of poor readers were good decoders but were relatively weak in analyzing the context of the sentence in reading. The performance of these children in the identification judgment tests suggested that they were aware of basic syntactic structures and could use them for perception of speech. However, they were inferior to good readers in using those structures productively as suggested by their relatively worse performance in the correction test. Thus, our data set limits to previous assertions that poor reading is not related to syntactical impairment (Gleitman & Rozin, 1977; Liberman, 1971; Mattingly, 1972; Shankweiler & Crain, 1986).

Of course, we do not claim to have found a causal relationship between syntactic ability and reading disorders. What we have seen is that, at least in Hebrew, there are poor readers of normal intelligence who are good decoders. Their performance suggests that there are aspects of poor reading that are not accounted for by deficient phonological processing. Moreover, we have shown that this impairment is associated with deficiencies in linguistic ability, here exemplified in the syntactic domain.

REFERENCES

Becker, C. A., & Killion, T. H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*. 3, 389-401.

Bentin, S., & Frost, R. (1987). Processing lexical ambiguity and visual word recognition in a deep orthography. *Memory & Cognition*, *15*, 13-24.

Brady, S. A. (1986). Short-term memory, processing and reading ability. *Annals of Dyslexia*, *36*, 138-153.

Brady, S. A., Shankweiler, D., & Mann, V. A. (1983). Speech perception and memory coding in relation to reading ability. *Journal of Experimental Child Psychology*, *35*, 345-367.

Brittain, M. M. (1970). Inflectional performance and early reading achievement. *Reading Research Quarterly*, *6*, 34-48.

Byrne, B. (1981). Deficient syntactic control in poor readers: Is a weak phonetic memory code responsible? *Applied Psycholinguistics*, *2*, 201-212.

Chomsky, N. (1969). *Acquisition of syntax in children from 5 to 10*. Cambridge, MA: MIT Press.

Cromer, W., & Wiener, M. (1966). Idiosyncratic response patterns among good and poor readers. *Journal of Consulting Psychology*, *30*, 1-10.

Fowler, A. E. (1988). Grammaticality judgment and reading skill in grade 2. *Annals of Dyslexia*, *38*, 73-94.

Glass, A. L., & Perna, J. (1986). The role of syntax in reading disability. *Journal of Learning Disabilities*, *19*, 354-359.

Gleitman, L. R., & Rozin, P. (1977). The structure and acquisition of reading: Relation between orthography and the structured language. In A. S. Reber & D .L. Scarborough (Eds.), *Towards a Psychology of Reading: The Proceedings of the CUNY Conference* (pp. 1-54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldman, S. R. (1976). Reading skill and the minimum distance principle. A comparison of listening and reading comprehension. *Journal of Experimental Child Psychology*, *22*, 123-142.

Goodman, G. O., McClelland, J. L., & Gibbs, R. W. (1981). The role of syntactic context in word recognition. *Memory & Cognition 9*, 580-586.

Guthrie, J. T. (1973). Reading comprehension and syntactic response in good and poor readers. *Journal of Educational Psychology*, *65*, 294-299.

Lasnik, H., & Crain S, (1985). On the acquisition of pronominal reference. *Lingua*, *65*, 135-154.

Liberman, I. Y. (1971). Basic research in speech and lateralization of language: Some implications for reading disability. *Bulletin of the Orton Society*, *21*, 71-87.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fisher F. W. (1977). Phonetic segmentation and recording in the beginning readers. In A. S. Reber & D. L. Scarborough (Eds.), *Towards a Psychology of Reading: The Proceedings of the CUNY Conference* (pp. 207-225). Hillsdale, NJ: Erlbaum Associates.

Lukatela, G., Kostić, A., Feldman, L. B, & Turvey, M. (1983). Grammatical priming of inflected nouns. *Memory & Cognition*, *11*, 59-63.

Mann, V. A, Liberman, I. Y., & Shankweiler, D. (1980). Children's memory for sentences and word strings in relation to reading ability. *Memory & Cognition*, *8*, 329-335.

Massaro, D. W., Jones, R. D., Lipscomb, C., & Scholz, R. (1978). Role of prior knowledge on naming and lexical decisions with good and poor stimulus information. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 498-512.

Mattingly, I. G. (1972). Reading, the linguistic process, and linguistic awareness. In J. F. Kavanagh & I. G. Mattingly (Eds.),

Language by ear and by eye (pp. 133-148). Cambridge, MA: MIT Press.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Loci of contextual effects on word recognition. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 98-118). New York: Academic Press.

Newcomer, P., & Magee, P. (1977). The performance of learning (reading) disabled children on a test of spoken language. *The Reading Teacher, 30*, 896-900.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A., Goldman, S., & Hogaboam, T. (1979). Reading skill and the identification of words in discourse context. *Memory & Cognition, 7*, 273-282.

Perfetti, C. A., & Roth, S. (1981). Some of the interactive processes in reading and their role in reading skill. In A. Lesgold & C. A. Perfetti (Eds.), *Interactive processes in reading* (pp. 269-298). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rozin, P., & Gleitman, L. (1977). The structure and acquisition of reading II: The reading process and the acquisition of the alphabetic principle. In A. S. Reber & D. L. Scarborough (Eds.), *Towards a Psychology of Reading: The Proceedings of the CUNY Conference* (pp. 55-142). Hillsdale, NJ: Erlbaum Associates.

Schvaneveldt, R. W., Ackerman, B., & Semlear, T. (1977). The effect of semantic context on children's word recognition. *Child Development, 48*, 612-616.

Schwantes, F. M., Boesl, S. L., & Ritz, E. G. (1980). Children's use of context in word recognition: A psycholinguistic guessing game. *Child Development, 51*, 730-736.

Shankweiler, D., & Crain S. (1986). Language mechanisms and reading disorders: A modular approach. *Cognition, 24*, 139-168.

Smith, F. (1971). *Understanding reading*. New York: Holt, Rinehart, and Winston.

Stanovitch, K. E., & West, R. F. (1981). The effect of sentence context on ongoing word recognition: Tests of a two process theory. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 658-672.

Stanovitch, K. E., West, R. F., & Feeman, D. J. (1981). A longitudinal study of sentence context effects in second-grade children: Tests of an interactive compensatory model. *Journal of Experimental Child Psychology, 32*, 185-199.

Vellutino, F. R. (1979). *Dyslexia: Theory and research*. Cambridge, MA: MIT Press.

Vogel, S. A. (1974). Syntactic abilities in normal and dyslexic children. *Journal of Learning Disabilities, 7*, 103-109.

West, R. F., & Stanovitch, K. E . (1978). Automatic contextual facilitation in readers of three ages. *Child Development, 49*, 717-727.

West, R. F., & Stanovitch, K. E. (1986). Robust effects of syntactic structure on visual word processing. *Memory & Cognition, 14*, 104-112.

Wright, B., & Garret, M. (1984). Lexical decision in sentences: Effects of syntactic structure. *Memory & Cognition, 12*, 31- 45.

FOOTNOTES

* *Journal of Experimental Child Psychology*, in press.

† Also Department of Psychology and School of Education, The Hebrew University, Jerusalem, Israel.

†† Also Department of Education, University of Connecticut, Storrs.

Haskins Laboratories Status Report on Speech Research
1989, SR-99/100, 195-210

Effect of Emotional Valence in Infant Expressions upon Perceptual Asymmetries in Adult Viewers

Catherine T. Best† and Heidi Fréya Queen††

Research on normal and brain-damaged adults indicates that the cerebral hemispheres are specialized for emotional as well as cognitive functions. However, controversy remains over which pattern of cerebral organization best accounts for emotion perception: overall right hemisphere (RH) superiority; RH specialization for negative emotion but left hemisphere (LH) specialization for positive emotion; RH specialization for negative emotion but no asymmetry for positive; or RH specialization for avoidance-related emotions and LH specialization for approach-related ones. Most studies of normal adults support overall RH specialization for emotion perception. However, there is some suggestion of valence effects in perceptual asymmetries, which may depend on engagement of emotional responses in the viewer. The present research examined asymmetries in adults' perception of smiling and crying infant expressions because, according to ethological theory, infant characteristics elicit heightened emotional responses in adults. Results from Experiment 1 supported RH specialization for perception of negative expressions, with a lack of asymmetry for positive expressions. Experiments 2 and 3 investigated whether this negative-valence effect can be attributed to differences in the involvement of LH feature-oriented versus RH holistic approaches to basic information processing, or rather to some other independent, emotion-related specialization of the hemispheres. The results of the latter two experiments were inconsistent with the information processing explanation. Discussion concludes with a suggestion that the adult's specialized RH sensitivity to infant crying expressions may have evolved in response to selection pressures for rapid response to signals that indicate potential threats to infant survival.

Research findings with both unilateral brain-damaged patients and normal adults have led to general consensus that the human cerebral hemispheres are differentially involved in emotional, as well as cognitive, processes. However, the exact pattern of hemispheric involvement in emotions remains controversial. According to the most widely-held view, the right hemisphere (RH) dominates overall in the perception and expression of emotion, across both negative and positive emotional valence (e.g., Campbell, 1978; Chaurasia & Goswami, 1975; Gainotti, 1972, 1988; Hirschman & Safer, 1982; Ladavas, Umilta & Ricci-Bitti, 1980; Ley & Bryden, 1979, 1981; Safer, 1981; Strauss & Moscovitch, 1981). For convenience, that view will be referred to here as the *RH hypothesis*. The major counter-proposal has been that the right hemisphere predominates in perception and expression of negative emotions, the left in positive emotions, a view we will refer to as the *valence hypothesis* (e.g., Ahern & Schwartz, 1979; Dimond & Farrington, 1977; Natale, Gur & Gur, 1983; Reuter-Lorenz & Davidson, 1981; Reuter-Lorenz, Givis & Moscovitch, 183; Rossi & Rosadini, 1967; Sackeim, Greenberg, Weiman,

This research was supported in part by NIH grant NS-24655 and by a Biomedical Research Support Grant from Wesleyan University to the first author, as well as by NIH grant HD-01994 to Haskins Laboratories.

We thank the following people for their help in completing the paper and the research described here: Christine Blackwell, Shama Chaiken, Linda Creem, Angelina Diaz, Glenn Feitelson, Sari Kalin, Dara Lee, Roxanne Shelton, Alicia Sisk, Leslie Turner and Jennifer K. Wilson for their help with stimulus preparation and with collecting and scoring data; Nathan Brody and Michael Studdert-Kennedy for their thoughtful, constructive comments on earlier drafts of the manuscript; and Katherine Hildebrandt (now Karraker) for making her infant photographs and emotional rating data available to us.

Gur, Hungerbuhler, & Geschwind, 1982; Silberman & Weingartner, 1986; Terzian, 1964). Several variations on the valence hypothesis have also been offered. Some evidence suggests that while negative emotions show differential right hemisphere involvement, there may be less hemispheric asymmetry for positive emotions (e.g., Dimond, Farrington & Johnson, 1976; Ehrlichman, 1988; Sackeim & Gur, 1978, 1980); we will call this view the *negative-valence hypothesis*. Another possibility is that the differential involvement of the left and right hemispheres in emotions may depend on the motivational qualities of approach versus avoidance, respectively, rather than on the positive versus negative valence of the emotion per se (e.g., Kinsbourne, 1978). According to Davidson and colleagues, such an approach-avoidance distinction between hemispheres pertains only to the subject's emotional experience (internal feeling-state) and expression (mediated by frontal lobes), but not to perception of emotions (parietal lobes), which show an overall right hemisphere superiority (Davidson, 1984; Davidson & Fox, 1982; Davidson, Schwartz, Saron, Bennett, & Goleman, 1979; Fox & Davidson, 1986, 1987, 1988). For brevity, we will call the latter proposal the *motivational hypothesis*.

This report focuses on asymmetries in normal adults for perception of infant facial emotional expressions. The majority of findings on perception of *adult* facial expressions by normal, neurologically-intact subjects have favored the RH hypothesis (e.g., Brody, Goodman, Halm, Krinzman & Sebrechts, 1987; Bryden, 1982; Bryden & Ley, 1983; Campbell, 1978; Carlson & Harris, 1985; Gage & Safer, 1985; Heller & Levy, 1981; Hirschman & Safer, 1982; Levy, Heller, Banich, & Burton, 1983; Ley & Bryden, 1979, 1981; Moscovitch, 1983; Safer, 1981; Segalowitz, 1985; Strauss & Moscovitch, 1981). These studies have typically found a left visual field (LVF) advantage, implying RH superiority, in perception of both positive and negative emotional expressions.

Only a few perceptual asymmetry studies have supported the valence hypothesis or its variants. In favor of the valence hypothesis, adults rate tachistoscopically-presented facial expressions more negatively when they are presented to the LVF-RH, but rate them more positively in the right visual field (RVF)-left hemisphere (LH), although the RH is better overall at differentiating among categories of emotion (Natale et al., 1983). Similarly, when subjects must identify the visual field containing an emotional expression during simultaneous tachistoscopic presentations of an emotional expression in one visual field and a neutral expression in the opposite visual field, they detect negative expressions more rapidly in the LVF-RH but detect positive expressions more rapidly in the RVF-LH (Reuter-Lorenz & Davidson, 1981; Reuter-Lorenz et al., 1983). The motivational hypothesis is supported by research on EEG responses in subjects viewing films of negative versus positive emotional expressions. Both adults (Davidson et al., 1979) and infants (Davidson & Fox, 1982) showed greater electrocortical activation of the right frontal lobe while viewing emotionally negative films, but greater activation of the left frontal lobe during positive films. However, parietal lobe activation was greater on the right than the left side at both ages for both types of film. No studies on perception of facial expressions support the negative-valence hypothesis. However, when emotionally negative films have been presented to a single hemisphere by having subjects wear half-silvered contact lenses (Dimond et al., 1976), or when emotionally negative odorants have been restricted to a single hemisphere via presentation to the ipsilateral nostril (Ehrlichman, 1988), subjects rated the stimuli presented to the RH as more intensely negative, without showing significant asymmetries for rating emotionally positive stimuli.

Why are there such inconsistent findings across studies of cerebral asymmetries in normal adults' perception of facial expressions? To some extent, they may be explained by variations in methodology and task requirements. Studies supporting the RH hypothesis have typically required subjects to recognize or discriminate facial expressions. In contrast, the studies favoring variants of the valence hypothesis have called for judgments about the emotionality of the stimuli. Recognition and discrimination judgments call upon basic cognitive-perceptual skills, and may be carried out by so-called "cold" cognitive abilities, whereas judgments about emotionality may more directly require the viewer to tap into emotional processes. This observation raises the related possibility that the presence or absence of a valence effect in perception may depend on the viewer's own *emotional response* to the stimuli, as Davidson (1984) and Ehrlichman (1988) have suggested. The viewer's emotional response, in turn, could very likely be influenced by whether the emotions represented in the

stimulus photographs are genuine or spontaneous versus simulated or posed. Clinical evidence indicates that *productions* of spontaneous and posed expressions are mediated by different neural pathways; the former is disturbed by damage to the temporal lobes or extrapyramidal system, while the latter is disturbed by frontal or pyramidal damage (e.g., Monrad-Krohn, 1924; Remillard, Anderman, Rhi-Sausi & Robbins, 1977; Rinn, 1984). These two classes of expression could therefore be expected to provide different information to the viewer about the emitter's internal emotional state. Perceivers would presumably be more likely to have an emotional response themselves while viewing genuine emotional expressions than they would while viewing simulated expressions. In this context, it is important to note that nearly all the data on asymmetries in perception of facial expressions have been obtained with stimuli containing posed rather than spontaneous emotional expressions.

For these reasons, we conducted a series of experiments requiring judgments about the emotionality of facial expression stimuli that should be more likely to elicit emotional responses in the viewer. Photographs of smiling and crying infants were chosen as the stimulus materials, based on several considerations. First, infant expressions are certainly more spontaneous and genuine, thus providing a more direct window on the infant's actual affective state, than are most adult expressions, especially those facial expressions that occur in social situations. Even in the case of so-called spontaneous expressions in adults, the facial display is often influenced to some extent by the forces of social conditioning and cultural display rules (Buck, 1986; Ekman, 1972), which would have much less or no influence on infants' expressions (e.g., Campos, Barrett, Lamb, Goldsmith, & Stenberg, 1983; Rothbart & Posner, 1985). It is generally assumed that infants do not begin to simulate, mask, or deliberately control their facial expressions until the second year of life (e.g., Campos et al., 1983; Oster & Ekman, 1978; Rothbart & Posner, 1985; Sroufe, 1979; cf. Fox & Davidson, 1988). Second, whereas adult expressions often involve complex mixtures of emotions, infant expressions tend to be simpler, reflecting purer examples of the basic categories of emotion (Campos et al., 1983; Izard, 1979; Izard, Huebner, Risser, McGinnes, & Dougherty, 1980). This characteristic of infant expressions may elicit simpler, more straightforward emotional responses from viewers than do more complex adult expressions. Third, ethological theory and research argue that infant expressions and appearance strongly tend to elicit emotional responses from adults, and do so to a greater extent than do the expressions and appearance of (unfamiliar) adults (e.g., Bowlby, 1969; Eibl-Eibesfeldt, 1975; Lorenz, 1935, 1981; Lorenz & Leyhausen, 1973). Adults' emotional responses to infant signals, including infant emotional expressions, are part of a mutually adapted system of evolved behaviors that promote the development of the relatively helpless human infant, which thus fosters the reproductive fitness of individuals displaying these characteristics, and ultimately the survival of the species. These responses to infants are, of course, particularly strong in their caregivers, but are present in all humans.

Infant crying and smiling are of particular interest to the present research when considered in ethological terms. Both serve to promote physical proximity between the infant and the caregiver, although for different reasons (e.g., Bowlby, 1969; Campos et al. 1983; Emde, Gaensbauer & Harmon, 1976). Smiling indicates a *positive* affective state and emotional *approach* of the infant toward the adult with which it is interacting. Infant smiling typically also elicits *positive* feelings and a corresponding *approach* response from that adult. The motivational tendencies associated with infant crying, however, differ for the infant and the responding adult. The infant's distress indicates *negative* feelings associated with some noxious stimulation or situation, and therefore a tendency for *withdrawal* in the infant. However, given an infant who cannot actively, physically withdraw itself, the crying typically elicits *negative* or distress feelings in nearby adults, which usually leads them (particularly caregivers) to attempt to aid the infant by eliminating the source of the distress. Thus, *approach* behavior is elicited in the adult. Based on this analysis of adults' responses to infant smiles and cries, the four hypotheses outlined above regarding cerebral organization for emotional processes would each predict a different pattern of asymmetries in adults' perception of infant emotional expressions. The RH hypothesis would predict an overall LVF advantage that is unaffected by the valence of the emotion expression. The valence hypothesis would predict a LVF-RH advantage for perception of infant crying expressions, but a RVF-LH advantage for perception of infant smiles. The negative-valence hypothesis would instead predict a LVF-RH advantage for crying expressions but a smaller or

nonexistent bias for perception of smiles. Finally, the motivational hypothesis would most likely predict a *RVF-LH* advantage for perception of both infant cries and smiles, because both should elicit approach responses in adult viewers. The RVF-LH bias might be larger for smiles than for cries, to the extent that smiles might elicit a stronger approach tendency when the adult viewers are not caregivers of the infants depicted.

Experiment 1 investigated these possibilities, using a set of photographs of infants' smiling and crying expressions. For this purpose, we employed the free-field viewing procedure developed by Levy, Heller, Banich and Burton (1983), in which subjects must choose which member of a pair of mixed-expression chimeras for each of a number of posers appears to be emotionally more intense (happier or sadder). Each pair of chimeras displays a half-neutral, half-emotional facial expression of a given poser; in one chimera, the emotional expression appears on the left side of the photograph, while in the other it appears on the right side. Significant asymmetrical biases in the viewers' choices between these pairs of chimeras are interpreted as reflecting asymmetrical activation of the cerebral hemispheres in response to the task, following Kinsbourne's (1978) proposal that activation of a one cerebral hemisphere will cause an attentional bias (increased perceptual sensitivity) favoring the contralateral spatial hemifield. Thus, if chimeras with the emotional expression on the left side of the photograph are perceived to be more emotional (happier or sadder) than those with the emotional expression on the right side, a LVF bias would be indicated, implying greater activation of the RH during the emotion judgment task. The converse pattern would indicate a RVF-LH bias (cf. Grega, Sackeim, Sanchez, Cohen & Hough, 1988). Levy et al. (1983) found a LVF-RH bias in adults' perception of half-neutral, half-smiling chimeras of adult posers' faces.

Experiments 2 and 3 were designed to examine whether the valence effect that was found in Experiment 1 could be attributed to differences in the extent to which judgments about smiling versus crying expressions involve the basic information processing approaches of the right versus the left hemisphere. The left hemisphere has been characterized as having a feature-oriented, analytical approach to information processing, whereas the right hemisphere approach has been described as holistic, gestalt-like, or synthetic (e.g., Bradshaw & Nettleton, 1981; Bryden, 1982; Levy, 1974). An effect of valence on asymmetries in the perception of emotional expressions might result from greater involvement of LH feature-oriented processing for one category of expression than for the other (cf. Moscovitch, 1983). For example, smiles may be perceived by simply focusing on whether the corners of the mouth are upturned or not, but perception of crying expressions may require the perceiver to attend to the overall configuration of mouth, eyes and brows. If so, then manipulating the stimulus properties to emphasize a focus on specific features should shift the degree and direction of visual field bias for judgments about negative and positive expressions. Alternatively, the valence effect could reflect differences in hemispheric specialization for positive and negative emotions per se, independent of the basic information processing skills of the two hemispheres. In the latter case, feature-oriented stimulus manipulations would not be expected to influence the valence effect on perceptual asymmetries. These possibilities were explored in the last two experiments reported here.

EXPERIMENT 1

Method

Subjects

Forty-six university students (23 female, 23 male) were included in this study; all had also participated in a related study of asymmetries in infants' facial expressions (Best & Queen, 1989). All were familial right-handers, a population that is more consistently and more strongly lateralized than non-right-handers for various hemispherically-specialized functions, including the perception of emotional expressions (Chaurasia & Goswami, 1975; Heller & Levy, 1981). Subjects completed a handedness checklist that assessed degree of hand preference on 10 common unimanual activities, including writing, as well as for the writing hand preference of immediate family members. To be considered strongly right-handed, subjects had to indicate a "strong" to "moderate" right-hand preference for all items, without ever switching hand preference during childhood, and both of their parents had to be right-handed. Four additional subjects were also tested but later eliminated for failure to meet the handedness criteria. All subjects had normal or corrected vision. They received $4.00 for their participation in the 40 minute test session.

Stimuli

The stimulus materials were generated from photographs of facial expressions by 10 normal,

full-term 7- to 13-month-old infants, which were originally taken by a portrait photographer for a series of studies on infant attractiveness (Hildebrandt & Fitzgerald, 1978, 1979, 1981). They were the same original photographs as those used in the Best and Queen (1989) study. Each infant provided a neutral facial expression and either a clearly negative (i.e., crying) or a clearly positive (i.e., smiling) expression, according to ratings obtained in an independent study (Hildebrandt, 1983). Four infants had crying expressions; the other 6 had smiling expressions. All photographs were of full-frontal facial views.

Two black-and-white 5 x 7 inch prints were made of each infant's neutral expression, along with two prints of the infant's smiling or crying expression. For each pair, one photograph was printed in normal orientation, and the other in mirror-reversed orientation. These were used to construct four mixed-expression chimeras for each infant (see Heller & Levy, 1981). Each print was cut down the exact facial midline, defined as the line connecting the point midway between the internal canthi of the eyes and the point in the center of the philtrum just above the upper lip. The two normal orientation chimeras for a given infant were made by joining the left half of the normal orientation print of the emotional expression (smile or cry) with the right half of the normal orientation neutral expression, and by joining the right half of the emotional expression with the left half of the neutral expression. For each chimera, the midlines of the hemifaces were aligned at the eyes and nose (the mouths often could not be exactly aligned because of differing degrees of mouth opening; see also Heller & Levy, 1981), and glued to a backing sheet. The mirror-reversed chimeras were constructed in like manner from the mirror-reversed prints of the emotional and neutral expressions. Comparison of results with the normal orientation chimeras versus the mirror-reversed chimeras allowed for assessment of any influence that infant expressive asymmetries might have upon the adults' responses.

Before reproduction, each chimera was centered behind an oval-shaped mattboard opening the size of the average photographed face, in order to screen out variations in facial outline and hair among the infants. Copies were made on a high-quality Kodak photocopier, using a gray-scale correction template that produces good resolution of photographic images. Each page contained either the pair of normal orientation chimeras for a given infant, or the pair of mirror-reversed chimeras for an infant, appearing one above the other. We used above-below rather than side-by-side pairings to avoid having subjects' choices of the more emotional chimera be confounded by hemispatial field biases resulting from the asymmetrical hemispheric activation that would be expected for such a laterally-specialized function (Kinsbourne, 1978). Each of the 10 infants was represented on four pages: 1) a normal-orientation pair in which the emotional expression was on the right side in the top picture; 2) a normal-orientation pair in which the emotional expression was on the left side in the top picture; 3) and 4) the pairs of mirror-reversed chimeras positioned as in items 1 and 2, respectively. Thus, there were 40 pages of paired chimeras. Test booklets were constructed with these pages ordered pseudorandomly, such that there were no more than 3 consecutive smiling infants or 3 consecutive crying infants, and no consecutive presentations of the same poser. At the top of each page one of the following questions was printed: "Which infant looks happier?" (for smiling-neutral chimeras) or "Which infant looks sadder?" (for crying-neutral chimeras).

Procedure

Subjects were tested in groups of 5-15 in a quiet room. They sat at separate desks and each had a copy of the test booklet, with a cover sheet of instructions. They circled "TOP" or "BOTTOM" on each numbered line (1-40) of an answer sheet to indicate which member of each pair of chimeras looked happier or sadder. Subjects proceeded through the booklet at their own pace, one page at a time without turning back or directly comparing pages.

Results

The data were converted to laterality ratios according to the formula (R-L)/(R+L), in which R = percent of chimera choices with the emotional expression on the right side of the picture (i.e., RVF preference), L = percent choices with the emotional expression on the left side (LVF preference), and R+L = 100%.[1] The laterality ratios thus range from -1.0 (extreme LVF bias) to +1.0 (extreme RVF bias). These values were then entered into a 2 x 2 analysis of variance (ANOVA), with the factors of emotion (cry, smile) and orientation of the photographs used in the chimera (normal, mirror-reversed). To determine whether the mean laterality ratios for each cell, and overall, showed a significant perceptual asymmetry (i.e., differed significantly from a score

of 0 laterality), *t*-tests were conducted. The alpha level correction for multiple *t*-tests (α = .05/number of *t*-tests) required a significance level of $p < .007$ for the latter tests.

There was a significant LVF bias overall (see Table 1 for all mean laterality ratios and *t*-tests).

Table 1. *Laterality ratios for the perception of mixed-expression chimeras of smiling and crying infants in normal and mirror-reversed orientation, Experiment 1.*

	Summary Statistics		
	Laterality Ratio[a]	*t* value[b]	*p*
Effects			
Overall perceptual bias	-.13	-4.78	.003
Emotion effect			
Smile	-.07	-2.57	ns
Cry	-.19	-4.84	.0000
Orientation effect			
Normal orientation	-.57	-17.84	.0000
Mirror-reversed	+.31	8.02	.0000
Emotion x Orientation interaction			
Smile			
Normal orientation	-.72	-22.07	.0000
Mirror-reversed	+.59	15.22	.0000
Cry			
Normal orientation	-.42	-7.84	.0000
Mirror-reversed	+.03	.54	ns

[a]Computed as [R-L]/[R+L], where R = percent choices with emotional expression on right of chimera, L = percent choices with emotional expression on left, and R + L = 100%. Negative scores indicate a left visual field bias, positive scores a right field bias.

[b]One-sample *t*-tests of whether the mean laterality ratio was significantly greater than 0, indicating a significant perceptual asymmetry.

However, a significant main effect of emotion, $F(1,45) = 10.09$, $p < .003$, indicated that the valence of the infant expressions influenced the degree of asymmetry in the adults' perception of the emotionality of the chimeras. Specifically, the LVF bias was significant for judgments of crying infants, but not for smiling infants. In addition, the orientation effect, $F(1,45) = 366.68$, $p = .0000$, revealed that the adult subjects' perceptual biases were quite sensitive to asymmetries in the infants' expressions themselves. The normal orientation chimeras, in which the infants' more expressive right hemiface (Best & Queen, 1989) appeared in the viewers' LVF, yielded a significant LVF bias. In contrast, the mirror-reversed chimeras, in which the infants' right hemiface appeared in the RVF, yielded a smaller but significant RVF bias. Finally, there was a significant Emotion x Orientation interaction, $F(1,45) = 66.80$, $p = .0000$ (see Table 1). For the smiling infants, there was a large difference in laterality ratios between the normal orientation chimeras, which showed a strong LVF bias, and the mirror-reversed chimeras, which showed a strong RVF bias. The viewers' perceptual asymmetries were less strongly influenced by orientation of the crying infant photos, showing a more moderate LVF bias for normal orientation chimeras and a very small, nonsignificant RVF bias for mirror-reversed chimeras. Simple effects tests of the interaction indicated that the orientation effect was significant, nonetheless, for both the crying expressions, $F(1,45) = 28.25$, $p = .0000$, and the smiling expressions, $F(1,45) = 703.32$, $p = .0000$. Furthermore, the emotion effect was significant for both normal orientation chimeras, $F(1,45) = 24.14$, $p = .0000$, and mirror-reversed chimeras, $F(1,45) = 63.29$, $p = .0000$.

DISCUSSION

Consistent with the hypothesis that cerebral asymmetries in emotional processes are influenced by emotional valence, the viewers in Experiment 1 showed a significant LVF bias in perception of negative but not positive infant emotional expressions. These results are compatible with the negative-valence hypothesis of cerebral organization for emotional processes (e.g., Ehrlichman, 1988). They do not as strongly support the valence hypothesis (e.g., Silberman & Weingartner, 1986; Tucker, 1981), because the LVF-RH bias in perception of negative infant expressions was not complemented by a RVF-LH bias in perception of positive expressions. The results also fail to support the motivational hypothesis of emotion lateralization (e.g., Davidson, 1984) in that infant cries and smiles did not yield a RVF-LH advantage, although both types of infant expression would be expected to elicit approach responses from adult viewers. The results also stand in contrast to the overwhelming majority of studies favoring the RH hypothesis, which have found a LVF advantage in adults' perception of adult facial expressions that is unaffected by valence (e.g., Bryden, 1982; Bryden & Ley, 1983). Only three studies on the perception of adult facial expressions have found a valence effect in neurologically intact adults, and these both involved tachistoscopic presentations

(Natale, Gur, & Gur, 1983; Reuter-Lorenz & Davidson, 1981; Reuter-Lorenz, Givis & Moscovitch, 1983). Thus it is remarkable that the perception of infant expressions showed a consistent valence effect even with the free-field task used here.

It was suggested earlier that a valence effect should be optimized by the heightened emotional response that infant expressions elicit from adults according to ethological theory. Indeed, informal observations revealed that many of the subjects smiled or showed other positive emotional responses to the infant faces during the test, whereas none of them showed such responses while completing a similar test with chimeric adult expressions during the same session. The suggestion of heightened sensitivity to infant emotional expressions is further corroborated by the finding that the perceptual responses of our adult subjects were strongly influenced by the infants' own expressive asymmetries. In contrast, Levy et al. (1983) failed to find significant evidence of perceivers' sensitivity to the expressive asymmetries of their adult posers in a similar free-field study using adult chimeric expressions.[2] In the present experiment, the viewers' sensitivity to asymmetrical information in the infant expressions was great enough to reverse their perceptual field asymmetries from a LVF bias when the more expressive right hemiface of the infants was on the left side of the chimera, to a smaller but significant RVF bias when the infants' right hemiface was on the right side of the chimera.

The interaction between the valence of the infants' emotional expressions and the orientation of the chimeras indicates that the relative impact of the viewers' perceptual biases and the infants' expressive asymmetries differed between judgments of negative and positive expressions. The pattern of this interaction provides additional support for the negative-valence effect. Although the left/right position of the infants' more expressive right hemiface within the chimeras influenced perception of both types of expressions, it was a much weaker determinant of judgments about cries than about smiles. That is, infant expressive asymmetries had relatively less influence, and the viewers' LVF-RH bias relatively more influence, in the adults' responses to cries than to smiles.

Thus, Experiment 1 provided clear evidence of a negative-valence effect on asymmetries in adults' perception of infant emotional expressions. However, it did not elucidate the underlying perceptual processes responsible for the phenomenon. As suggested in the introduction, one possible source of the effect might be that negative expressions are perceived in terms of the configuration of the whole face (i.e., the gestalt of the features within the "frame" of a face outline and hair), whereas perception of positive expressions may focus upon the mouth as a single distinguishing feature (Moscovitch, 1983). The former approach would call more heavily upon the holistic abilities of the right hemisphere, while the latter approach would be more suited to the feature-oriented analytic abilities of the left hemisphere (e.g., Bradshaw & Nettleton, 1981; Levy, 1974). If the influence of emotional valence is attributable to such differences in the perceptual approach to crying and smiling expressions, then the negative-valence effect, and indeed the overall LVF bias, should become attenuated when the viewers' attention is progressively restricted to narrower sources of emotional information in the infants' faces, such as the patterning of the central facial features without the contextual "frame" of the facial outline, hair, and other peripheral details. This manipulation should lead subjects to use a more feature-oriented, analytic approach, and thus to rely more heavily on left hemisphere information processing strategies. Alternatively, it may be that the viewers' actual emotional responses to crying and smiling infants, rather than the information processing strategy, are responsible for the valence effect. If so, then the negative-valence effect and the overall perceptual field bias should appear even when the viewer's attention is focused away from the gestalt of the faces and toward subcomponents or specific features. The next two experiments were designed to systematically examine these possibilities.

EXPERIMENT 2

If the holistic or gestaltlike perceptual specialization attributed to the right hemisphere is responsible for the finding of a LVF for crying but not smiling expressions, we would expect that perception of cries focuses on the gestalt of the whole face, i.e., the patterning of facial features within the configurational "frame" of the face. If the crying expressions evoked a greater degree of right hemisphere involvement because they were perceived in a more holistic manner, then removal of the peripheral configurational information such the facial outline should attenuate or eliminate the negative-valence effect. If, on the other hand, the valence effect derives from the emotional

nature of the response to the infant expressions, then it should persist even for judgments of the central facial features alone.

To restrict the viewers' attention to the patterning of the central features of the eyes/brows, mouth and nose, we deleted the unwanted peripheral details (e.g., face outline, hair, cheeks) from optically-digitized versions of the original photographs via computer. A new group of subjects made choices between each pair of the mixed-expression chimeras generated from these computer-edited infant expressions, as in Experiment 1.

Method

Subjects

Ninety-six familial right-handed university and high school students (51 female, 45 male) participated in Experiment 2. All had normal or corrected vision, and all had participated in the Best and Queen (1989) study. The university students received $4.00 for their participation in the 45 minute session; the high school students were unpaid volunteers.

Stimuli

High-quality photocopies of the original photographs from Experiment 1 were computer-digitized and edited, using an Apple® Macintosh™ 512+ computer (see Best and Queen, 1989, for details). The cheeks, ears, chin, hair, and face outline were removed from the digitized pictures, and the resulting edited images were printed in both normal and mirror-reversed orientation. These were then used to generate mixed-expression chimeras of each infant (see examples, Figure 1), which were assembled into a 40-page test booklet and reproduced with a high-quality photocopier, as in Experiment 1.

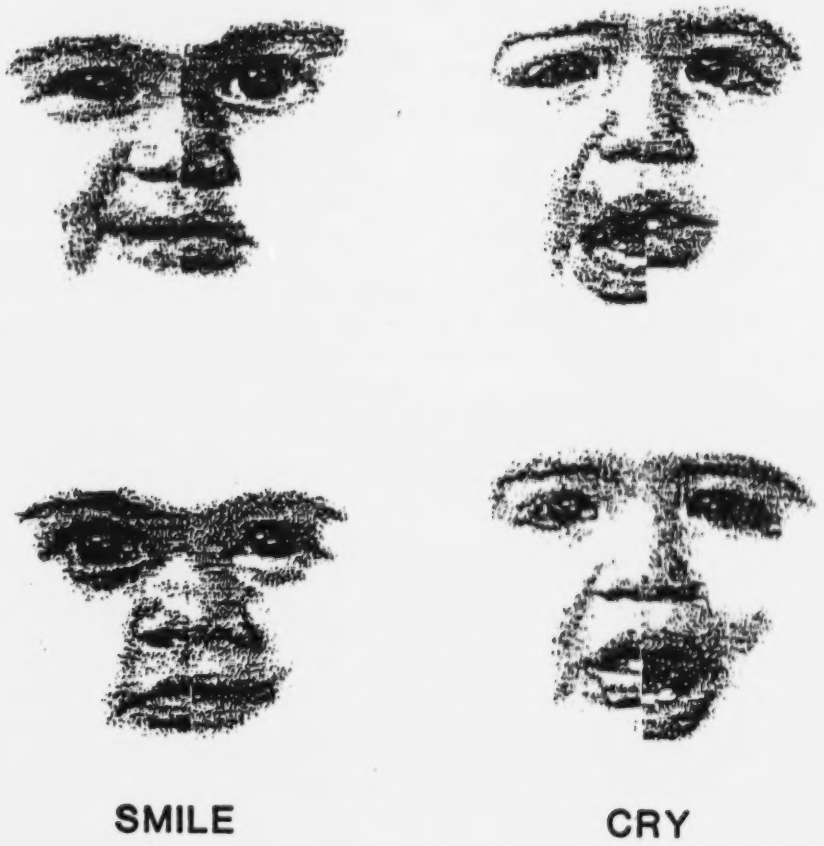


Figure 1. Examples of digitized mixed-expression chimeras of a smiling and a crying infant, with the emotional expressions on the left versus the right side of the picture, Experiment 2.

Procedure

Subjects completed the test booklet under the same conditions and instructions as in Experiment 1.

Results

The data were transformed to laterality ratios and analyzed as in Experiment 1. The significance level for the multiple *t*-tests was set at $p <= .007$, as before.

Again, there was a significant overall LVF bias in perception of the emotional chimeras (see Table 2 for mean laterality ratios and *t*-tests). The magnitude of this LVF bias did not change significantly from that found in Experiment 1. The emotion effect was significant, as in Experiment 1, $F(1,95) = 7.76$, $p < .007$, again indicating a stronger LVF bias for crying expressions than for smiling expressions. The magnitude of the difference in visual field biases between the crying and the smiling infants did not differ significantly from those found in Experiment 1, according to *t*-test. The orientation effect was also significant, $F(1,95) = 432.01$, $p = .0000$, indicating that the infant's expressive asymmetries affected the viewers' judgments. There was a significant LVF bias when the infants' more expressive right hemiface was on the left side of the chimeras, but a RVF bias when it was on the right side. The Emotion x Orientation interaction was also significant, $F(1,95) = 63.64$, $p = .0000$, following the pattern found in Experiment 1. For the smiling infants, the normal orientation chimeras showed a strong LVF bias, and the mirror-reversed chimeras showed a strong RVF bias (see Table 2). The orientation of the crying infant photos showed a similar but weaker influence, yielding a moderate LVF bias for normal orientation chimeras, and a much smaller RVF bias for mirror-reversed chimeras. According to simple effects tests of this interaction, nevertheless, the orientation effect was significant for both the crying expressions, $F(1,95) = 13.23$, $p < .0005$, and the smiling expressions, $F(1,95) = 65.76$, $p = .0000$, and the emotion effect was significant for both normal orientation chimeras, $F(1,95) = 559.26$, $p = .0000$, and mirror-reversed chimeras, $F(1,95) = 104.62$, $p = .0000$.

Table 2. *Laterality ratios for the perception of mixed-expression chimeras of smiling and crying infants in normal and mirror-reversed orientation, Experiment 2.*

	Summary Statistics		
	Laterality Ratio[a]	*t* value[b]	*p*
Effects			
Overall perceptual bias	-.11	-5.33	.0000
Emotion effect			
Smile	-.07	-3.10	.003
Cry	-.15	-5.62	.0000
Orientation effect			
Normal orientation	-.49	-19.13	.0000
Mirror-reversed	+.26	8.83	.0000
Emotion x Orientation interaction			
Smile			
Normal orientation	-.56	-19.66	.0000
Mirror-reversed	+.42	12.34	.0000
Cry			
Normal orientation	-.41	-11.08	.0000
Mirror-reversed	+.11	2.94	.005

[a]Computed as [R-L]/[R+L], where R = percent choices with emotional expression on right of chimera, L = percent choices with emotional expression on left, and R + L = 100%. Negative scores indicate a left visual field bias, positive scores a right field bias.

[b]One-sample *t*-tests of whether the mean laterality ratio was significantly greater than 0, indicating a significant perceptual asymmetry.

DISCUSSION

The results of Experiment 2 replicated those of Experiment 1, even though the gestalt of the whole faces had been modified by removal of the facial outline and of extraneous details other than the pattern of the central facial features. In fact, the magnitude of the effects failed to differ significantly from those found in Experiment 1. These findings suggest that the viewers' perception of the more complete chimeric photographs in the previous study had been based upon the central facial features rather than on their relation to peripheral information such as the "frame" of the facial outline and hair. It also suggests that the negative-valence effect is due to the emotional nature of the perceptual response to the faces, rather than to differential involvement of the right hemisphere's putative holistic approach and the left hemisphere's putative feature-analytic approach to negative vs. positive expressions, respectively.

Perhaps, however, the stimulus manipulations of Experiment 2 did not provide sufficient interference with the gestalt of the facial expressions to disrupt the right hemisphere's greater holistic response to crying than to smiling expressions. A clearer disruption of the holistic approach would involve focusing the viewers' attention on even more narrowly-defined features of the faces, such as the mouth or the eye region, both of which would be expected to carry much of the information conveyed in an emotional expression. The third experiment investigated the possibility that such a narrow feature-oriented approach would influence the negative-valence effect.

EXPERIMENT 3

The purpose of the third experiment was to determine the effect of restricting the viewers' attention to the emotional expression displayed in single, isolated facial features, which should more definitely bias the perceptual approach toward the analytic, feature-oriented abilities ascribed to the left hemisphere. If information processing differences in the perception of smiles as compared to cries were responsible for the negative-valence effect in the earlier studies, then this manipulation should either eliminate the valence effect in perceptual asymmetries or shift it to a strong RVF bias for smiling expressions with a weak or nonexistent LVF bias for crying expressions. However, if the negative-valence effect arises instead from the emotional nature of the judgments, then the pattern of findings and the magnitude of the effects should be impervious to this manipulation.

In restricting viewers' attention to specific facial features, we focused on the expressive patterning of the mouth versus that of the the eyes. In a previous paper (Best & Queen, 1989), we had found that the infants' right hemiface bias in expressiveness was specific to the mouth region of the face, and was not present in the eye region, even though the viewers had no difficulty making emotionality judgments about pairs of chimeras generated from either of these isolated facial regions. Because cortical input to the mouth region is contralateral, whereas input to the eye region is bilateral, those earlier results had suggested that lateralized cortical specializations, rather than more peripheral factors, are responsible for the right hemiface bias in infant expressions. Thus, a second purpose of the present experiment was to test whether adults' perceptual asymmetries are influenced by the difference in asymmetrical patterning between the eye and the mouth regions of the infants' expressions. For Experiment 3, a new group of judges was presented with an "upper face" test and a "lower face" test that employed further modifications of the digitized, edited infant expressions developed in Experiment 2.

Method

Subjects

Fifty-four familial right-handed university students (27 female, 27 male) participated in this experiment. They received $4.00 for participating in a 45 minute session. All had normal or corrected vision, and all had participated in the Best and Queen study (1989).

Stimuli

The digitized, edited faces from Experiment 2 were again revised to produce an "upper face" test, for which all facial features other than the eyes, brows and bridge of the nose were removed , and a "lower face" test, for which all features other than the mouth and the tip of the nose were eliminated. Mixed-expression chimeras were generated separately for the eyes/brows and for the mouth (see Figure 2).[3] Two 40-page test booklets were constructed as in the previous experiments, one for the "lower face" test and one for the "upper face" test. These were duplicated on a high-quality photocopier, as before.

Figure 2. Examples of digitized mixed-expression chimeras of the eye region and the mouth region of a smiling and a crying infant, with the emotional expressions on the left versus the right side of the picture, Experiment 3.

Procedure

The subjects were tested under the same conditions and instructions as in the first two experiments. They each completed the "lower face" test first and the "upper face" test second. Pilot testing suggested that judgments about the eyes/brows might be more difficult than judgments about the mouth; this test order thus allowed more practice before the more difficult test.

Results

The data were handled as in Experiments 1 and 2, except that the ANOVA for this experiment included a third, new factor: face part (mouth vs. eyes), and the alpha level adjustment for multiple *t*-tests set the significance level at $p <= .002$.

Once again, there was an overall LVF bias (see Table 3 for mean laterality ratios and *t*-tests). The magnitude of this LVF bias did not differ significantly from that found in either of the previous experiments. The emotion effect was again significant, $F(1,53) = 13.96$, $p < .0005$, indicating that the valence of the infant emotion continued to influence the viewers' perception of the expressions, even when their attention was restricted to isolated facial features. The crying infants elicited a significant LVF bias, which reversed to a nonsignificant RVF bias for the smiling infants. The magnitude of this valence effect again failed to differ significantly from that found in Experiments 1 and 2. As in the first two experiments, there was also a significant orientation effect, $F(1,53) = 39.70$, $p = .0000$. There was a LVF bias only for judgments of normal orientation expressions, when the infants' right hemiface was on the left side of the chimeras. The mirror-reversed chimeras elicited a small, nonsignificant RVF bias. The Emotion x Orientation interaction was significant as well, $F(1,53) = 8.97$, $p < .004$. Also as before, orientation had a smaller effect on perceptual asymmetries in response to crying than to smiling expressions. There was a LVF bias for crying infants, which was significant for the normal orientation chimeras, but not for mirror-reversed chimeras. In contrast, the normal orientation smiling infants

evoked an even larger LVF bias, but the mirror-reversed smile expressions produced a significant RVF bias. According to simple effects tests of this interaction, the orientation effect was significant for both the crying expressions, $F(1,53) = 5.13$, $p = .02$, and the smiling expressions, $F(1,53) = 76.97$, $p < .0000$. However, the emotion difference was significant only for the mirror-reversed chimeras, $F(1,53) = 21.28$, $p < .0000$, and not for the normal orientation chimeras.

Table 3. *Laterality ratios for the perception of mixed-expression chimeras of smiling and crying infants in normal and mirror-reversed orientation, eye versus mouth, Experiment 3.*

	Summary Statistics		
	Laterality Ratio[a]	*t* value[b]	*p*
Effects			
Overall perceptual bias	-.08	-4.44	.0000
Emotion effect			
Smile	-.02	-0.86	ns
Cry	-.14	-5.14	.0000
Orientation effect			
Normal orientation	-.19	-8.47	.0000
Mirror-reversed	+.03	1.09	ns
Emotion x Orientation interaction			
Smile			
Normal orientation	-.17	-6.99	.0000
Mirror-reversed	+.14	4.79	.0000
Cry			
Normal orientation	-.20	-5.62	.0000
Mirror-reversed	-.08	-1.85	ns
Face part x Orientation interaction			
Mouth region			
Normal orientation	-.35	-10.31	.0000
Mirror-reversed	+.20	4.93	.0000
Eye region			
Normal orientation	-.03	-1.10	ns
Mirror-reversed	-.14	-3.98	.0002
Face part x Orientation x Emotion interaction			
Mouth region			
Smile			
Normal orientation	-.57	-17.31	.0000
Mirror-reversed	+.52	11.92	.0000
Cry			
Normal orientation	-.13	-2.32	ns
Mirror-reversed	-.11	-1.93	ns
Eye region			
Smile			
Normal orientation	+.23	5.672	.0000
Mirror-reversed	-.24	-4.95	.0000
Cry			
Normal orientation	-.28	-6.23	.0000
Mirror-reversed	-.05	-0.97	ns

[a]Computed as [R-L]/[R+L], where R = percent choices with emotional expression on right of chimera, L = percent choices with emotional expression on left, and R + L = 100%. Negative scores indicate a left visual field bias, positive scores a right field bias.
[b]One-sample *t*-tests of whether the mean laterality ratio was significantly greater than 0, indicating a significant perceptual asymmetry.

The new face part factor also entered into two significant interactions. The Face part x Orientation interaction, $F(1,53) = 101.86$, $p = .0000$, indicated that the infants' expressive asymmetries had a greater influence on perception of the mouth than of the eyes. The normal orientation chimeras of the mouth yielded a large LVF bias in perception, while the mirror-reversed versions yielded a large RVF bias. In contrast, the eye chimeras produced a smaller but significant LVF bias when presented in *mirror-reversed* orientation, which became nonsignificant for the normal orientation eyes. Simple effects tests of the Face part x Orientation interaction showed that the face part difference was significant for both the normal orientation chimeras, $F(1,53) = 62.39$, $p = .0000$, and the mirror-reversed ones, $F(1,53) = 44.88$, $p < .0000$. Moreover, the orientation effect was significant both for the mouth, $F(1,35) = 119.58$, $p = .0000$, and for the eyes, $F(1,53) = 6.68$, $p < .01$.

The Face part x Orientation x Emotion interaction was also significant, $F(1,53) = 165.97$, $p = .0000$. The results for the mouth region chimeras followed the pattern found in the earlier experiments. The smiling mouths produced a large LVF bias for normal orientation chimeras, and a large RVF bias for mirror-reversed ones. Yet orientation had no appreciable effect on perception of the crying mouths, which showed a nonsignificant LVF bias for both normal orientation and mirror-reversed chimeras. The crying eyes yielded a LVF bias that was significant for normal orientation chimeras, but not for mirror-reversed ones, also consistent with the earlier findings. The smiling eyes, however, elicited a RVF bias for normal orientation chimeras, but a LVF bias for mirror-reversed chimeras. The direction of this orientation effect for judgments of smiling eyes was the opposite from the pattern of the Emotion x Orientation interactions found in Experiments 1 and 2, where the normal orientation was associated with LVF bias and the mirror-reversed orientation with RVF bias. Thus, the smiling eyes appear to be responsible for the Face part x Orientation interaction. Nonetheless, it is still clear that, for judgments of both the mouth and the eyes, orientation had a greater effect on perceptual responses toward the smiling expressions than toward the crying expressions. According to simple effects tests, the orientation effect was significant for crying eyes, $F(1,53) = 53.48$, $p = .0006$, for smiling mouths, $F(1,53) = 457.26$, $p = .0000$, and for smiling eyes, $F(1,53) = 11.06$, $p < .002$, but not for crying mouths. Also, the emotion effect was significant for eyes in normal orientation, $F(1,53) = 56.96$, $p = .0000$, and in mirror-reversed orientation, $F(1,53) = 9.72$, $p < .003$, as well as for mouths in normal orientation, $F(1,53) = 58.74$, $p = .0000$, and in mirror-reversed orientation, $F(1,53) = 90.83$, $p = .0000$.

DISCUSSION

The significant emotion effect was not diminished relative to the two earlier experiments, in spite of restricting the viewers' attention to isolated facial features. This result strongly suggests that the negative-valence effect on asymmetries in perception of infant emotional expressions derives from the emotional nature of the judgments rather than from differences in the information processing of negative versus positive expressions. Moreover, differences in perception of the eye and mouth regions suggest that the viewers were sensitive to differences in the expressive asymmetries displayed by those facial regions. Consistent with the Best & Queen (1989) finding that the right hemiface bias in infant expressions was significant only for the mouth region, the viewers in the present study were more affected by the orientation of the mouth than by the orientation of the eyes. For both the eyes and the mouth, however, the negative-valence effect held up (the Face part x Emotion interaction was not significant), in that there was a weaker orientation effect, or greater effect of perceptual asymmetry, for crying expressions than for smiling expressions. That is, the LVF bias in perception of infant emotional expressions was stronger (less affected by the infants' *expressive* asymmetries) for crying than for smiling expressions, suggesting greater right hemisphere involvement in perception of negative than in positive emotions.

GENERAL DISCUSSION

The results of all three experiments support the hypothesis that the right hemisphere is specialized for perception of negative emotion, but that perception of positive emotion is less strongly lateralized, a view we have referred to as the negative-valence hypothesis (e.g., Ehrlichman, 1988). The valence hypothesis of RH specialization for negative emotion, but LH specialization for positive emotion, was not supported, in that perception of infant smiles failed to show a significant RVF-LH bias. Support was also lacking for the motivational hypothesis, because the approach response that adults would be expected

to show toward both crying and smiling infants did not result in a RVF-LH perceptual bias, as predicted.

Given that the majority of studies on asymmetries in perception of adults' facial expressions have instead supported the RH hypothesis—that the right hemisphere is specialized for all emotion, regardless of valence—the present findings suggest that the influence of valence on perceptual asymmetries may depend on some involvement of emotional responses in the viewer. Our task, like those used in other reports that have supported variants of the valence hypothesis, called for judgments about the emotionality of the stimulus expressions. In contrast, discrimination or recognition judgments were required by many of the studies that obtained an overall RH advantage for emotion perception.

However, judgments about emotionality may not alone suffice to produce a negative-valence effect on perceptual asymmetries. Levy et al. (1983) presented subjects with pairs of mixed-expression chimeras of half-neutral, half-smiling adult faces in a free-field task and asked for judgments about the relative emotionality of each pair, as in the present study, yet those researchers obtained a significant LVF-RH advantage for perception of positive emotion. Our use of infant facial expressions to increase the likelihood of the viewers' emotional response to the stimuli may have been important in obtaining an influence of valence on perceptual asymmetries. This suggestion is corroborated by our finding that the subjects in Experiments 1 and 2 showed a strong, significant LVF bias in their perception of the Levy et al (1983) adult chimeric smiling expressions (Best & Queen, 1987), whereas they had instead shown a nonsignificant (Experiment 1) or small LVF bias (Experiment 2) in their perception of infant chimeric smiling expressions. Even stronger support is provided by Chaiken (1988), who used the same free-field chimeric face technique to compare adults' perceptual asymmetries for smiling and crying adult expressions versus smiling and crying infant expressions. Her subjects showed a significant valence effect in response to the infant expressions, but no valence effect in response to the adult expressions. It should be noted, however, that the viewers' actual emotional responses to the infant and adult stimuli were not directly assessed in any of these studies. Thus, further research is needed to test the hypothesis that emotional responses in the viewer are crucial in producing a valence effect on asymmetries in the perception of emotional expressions.

The results of experiments 2 and 3 also indicate that the negative-valence effect on perception of infant expressions cannot be explained by differences in the balance between the basic information processing approaches of the two hemispheres during the perception of negative versus positive emotions. Although stimulus manipulations designed to focus the viewers' attention on progressively more restricted features of the infant facial expressions might have shifted perception toward the putative analytical, feature-oriented approach of the LH, these manipulations did not influence the overall degree of asymmetry in emotion perception. Nor, more importantly, did the manipulations change the magnitude of the valence effect on perception. These findings thus suggest that the negative-valence effect in perception of infant emotional expressions reflects an aspect of hemispheric specialization that is independent of information processing asymmetries. The most likely basis for this separate characteristic of hemispheric specialization is an asymmetry in emotional responsiveness to stimuli, such that the RH shows greater sensitivity in the perception of negative emotion.

It is interesting to note that the adult's LVF-RH bias in perception of infants' crying expressions is compatible with recent findings that emotional expressions are more intense on the right side of the infant's face (Best & Queen, 1989; Rothbart, Taylor & Tucker, 1989). That is, in face-to-face interactions, the infant's more expressive hemiface would appear in the adult's more sensitive visual hemispatial field (see Introduction; also Kinsbourne, 1978), presumably increasing the likelihood of the adult's emotional response to the infant. This pattern of spatial compatibility between perceiver and producer of an emotional expression does not hold in the case of adults interacting face-to-face with other adults. Adults instead show a *left* hemiface bias in expressiveness, which would place the more expressive hemiface in the viewer's *less* sensitive hemispatial field. The enhancement of adults' sensitivity and responsiveness to infant expressions, relative to adult expressions, is consistent with ethological theory, as discussed in the general introduction. But why should there be greater compatibility between infant expressive asymmetry and adult perceptual asymmetry in the case of crying expressions than of smiling expressions? Perhaps this can be related to

differences in the imperativeness of adult responses to infant distress and pleasure states. Presumably, infant distress may indicate some possible danger to the infant, requiring immediate action on the part of the caregiver or other adult, whereas an infant's smile does not signal the need for such immediate action. Therefore, the evolutionary pressure for enhanced responsiveness to infant crying expressions would have been greater than the pressure for enhanced responsiveness to infant smiles.

REFERENCES

Ahern, G. L., & Schwartz, G. E. (1979). Differential lateralization for positive versus negative emotion. *Neuropsychologia, 17*, 693-698.

Best, C. T., & Queen, H. F. (1989). Baby, it's in your smile: Right hemiface bias in infant emotional expressions. *Developmental Psychology, 25*, 264-276.

Best, C. T., & Queen, H. F. (1987). [The free-field chimeric face technique is sensitive to asymmetries in adult posers' facial expressions]. Unpublished data.

Bowlby, J. (1969). *Attachment and loss* (Vol. 1). New York: Basic Books.

Bradshaw, J. L., & Nettleton, N. C. (1981). The nature of hemispheric specialization in man. *The Behavioral and Brain Sciences, 4*, 51-91.

Brody, N., Goodman, S. E., Halm, E., Krinzman, S., & Sebrechts, M. M. (1987). Lateralized affective priming of lateralized affectively valued target words. *Neuropsychologia, 25*, 935-946.

Bryden, M. P. (1982). *Laterality: Functional asymmetry in the intact brain*. New York: Academic Press.

Bryden, M. P. & Ley, R. G. (1983). Right hemispheric involvement in imagery and affect. In E. Perecman (Ed.), *Cognitive processing in the right hemisphere*. New York: Academic Press.

Bryden, M. P., & Sprott, D. A. (1981). Statistical determination of degree of laterality. *Neuropsychologia, 19*, 571-581.

Buck, R. (1986). The psychology of emotion. In J. E. LeDoux & W. Hirst (Eds.), *Mind and brain: Dialogues in cognitive neuroscience* (pp. 275-300). London, England: Cambridge University Press.

Campbell, R. (1978). Asymmetries in interpreting and expressing a posed facial expression. *Cortex, 19*, 327-342.

Campos, J. J., Barrett, K. C., Lamb, M. E., Goldsmith, H. H., & Stenberg, C. (1983). Socioemotional development. In P. H. Mussen (Ed.), M. M. Haith & J. J. Campos (Vol. Eds.), *Handbook of child development: Vol. II. Infancy and developmental psychobiology* (pp. 784-915). New York: John Wiley & Sons.

Carlson, D. F., & Harris, L. J. (1985). Perception of positive and negative emotion in free viewing of asymmetrical faces: Sex and handedness effects. Paper presented at the *International Neuropsychology Society* meeting, February 6-8, San Diego.

Chaiken, S. (1988). *A free vision test of developmental changes in hemispheric asymmetries for perception of infant and adult emotional expressions*. Unpublished M. A. thesis, Wesleyan University, Middletown CT.

Chaurasia, B. D., & Goswami, H. K. (1975). Functional asymmetry in the face. *Acta Anatomica, 91*, 154-160.

Davidson, R. J. (1984). Affect, cognition, and hemispheric specialization. In C. E. Izard, J. Kagan, & R. Zajonc (Eds.), *Emotion, cognition and behavior* (pp. 320-365). New York, NY: Cambridge University Press.

Davidson, R. J., & Fox, N. A. (1982). Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants. *Science, 218*, 1235-1237.

Davidson, R. J., Schwartz, G. E., Saron, C., Bennett, J., & Goleman, D. J. (1979). Frontal versus parietal EEG asymmetry during positive and negative affect. *Psychophysiology, 16*, 202-203.

Dimond, S. J., & Farrington, L. (1977). Emotional response to films shown to the right and left hemisphere of the brain measured by heart rate. *Acta Psychologica, 41*, 255-260.

Dimond, S. J., Farrington, L., & Johnson, P. (1976). Differing emotional response from right and left hemispheres. *Nature, 261*, 690-692.

Ehrlichman, H. (1988). Hemispheric asymmetry and positive-negative affect. In D. Ottoson (Ed.), *Duality and the unity of the brain*. London: MacMillan Press.

Eibl-Eibesfeldt, I. (1975). *Ethology: The biology of behavior*. New York: Holt, Rhinehart and Winston.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. *Nebraska symposium on motivation* (pp. 207-283). Lincoln: University of Nebraska Press.

Ekman, P., & Freisen, W. V. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.

Emde, R. N., Gaensbauer, T. J., & Harmon, R. J. (1976). *Emotional expression in infancy: A biobehavioral study*. New York: International Universities Press.

Fox, N. A., & Davidson, R. J. (1986). Taste-elicited changes in facial signs of emotion and the asymmetry of brain electrical activity in newborn infants. *Neuropsychologia, 24*, 417-422.

Fox, N. A., & Davidson, R. J. (1987). Electroencephalogram asymmetry in response to the approach of a stranger and maternal separation in 10-month-old infants. *Developmental Psychology, 23*, 233-240.

Fox, N. A., & Davidson, R. J. (1988). Patterns of brain electrical activity during facial signs of emotion in 10-month-olds. *Developmental Psychology, 24*, 230-236.

Gage, D. F., & Safer, M. A. (1985). Hemisphere differences in the mood state-dependent effect for recognition of emotional faces. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11*, 752-763.

Gainotti, G. (1972). Emotional behavior and hemispheric side of lesion. *Cortex, 8*, 41-55.

Gainotti, G. (1988). Disorders of emotional behavior and of autonomic arousal resulting from unilateral brain damage. In D. Ottoson (Ed.), *Duality and the unity of the brain*. London: MacMillan Press.

Grega, D. M., Sackeim, H. A., Sanchez, E., Cohen, B. H., & Hough, S. (1988). Perceiver bias in the processing of human faces: Neuropsychological mechanisms. *Cortex, 24*, 91-117.

Heller, W., & Levy, J. (1981). Perception and expression of emotion in right-handers and left-handers. *Neuropsychologia, 19*, 263-272.

Hildebrandt, K. A. (1983). Effect of facial expression variations on ratings of infants' physical attractiveness. *Developmental Psychology, 19*, 414-417.

Hildebrandt, K. A., & Fitzgerald, H. E. (1978). Adults' responses to infants varying in perceived cuteness. *Behavioural Processes, 3*, 159-172.

Hildebrandt, K. A., & Fitzgerald, H. E. (1979). Adults' perceptions of infant sex and cuteness. *Sex Roles, 5*, 471-481.

Hildebrandt, K. A., & Fitzgerald, H. E. (1981). Mothers' responses to infant physical appearance. *Infant Mental Health Journal, 2*, 56-61.

Hirschman, R. S., & Safer, M. A. (1982). Hemisphere differences in perceiving positive and negative emotions. *Cortex, 18*, 569-580.

Izard, C. (1979). *The maximally discriminative facial movement coding system*. Newark, DE: University of Delaware Press.

Izard, C., Huebner, R. R., Risser, D., McGinnes, G. C., & Dougherty, L. M. (1980). The young infant's ability to produce discrete emotional expressions. *Developmental Psychology, 16*, 132-140.

Kinsbourne, M. (1978). The biological determinants of functional bisymmetry and asymmetry. In M. Kinsbourne (Ed.), *Asymmetrical functions of the brain*. New York: Cambridge University Press.

Kuhn, G. M. (1973). The Phi coefficient as an index of ear differences in dichotic listening. *Cortex, 9*, 447-457.

Ladavas, E., Umilta, C., & Ricci-Bitti, P. E. (1980). Evidence for sex differences in right hemisphere dominance for emotions. *Neuropsychologia, 18*, 361-366.

Levy, J. (1974). Psychobiological implications of bilateral asymmetry. In S. J. Dimond & G. Beaumont (Eds.), *Hemisphere function in the human brain*. London: Elek Books.

Levy, J., Heller, W., Banich, M. T., & Burton, L. A. (1983). Asymmetry of perception in free viewing of chimeric faces. *Brain and Cognition, 2*, 404-419.

Ley, R. G, & Bryden, M. P. (1979). Hemispheric differences in recognizing faces and emotions. *Brain and Language, 7*, 127-138.

Ley, R. G., & Bryden, M. P. (1981). Consciousness, emotion, and the right hemisphere. In R. Stevens & G. Underwood (Eds.), *Aspects of consciousness: Structural issues*. New York: Academic Press.

Lorenz, K. Z. (1935). *Studies in animal and human behavior*.(Vol. 1). Cambridge, MA: Harvard University Press.

Lorenz, K. Z. (1981). *The foundations of ethology*. New York: Springer-Verlag.

Lorenz, K. Z., & Leyhausen, P. (1973). *Motivation of human and animal behavior*. New York: Van Nostrand.

Monrad-Krohn, G. H. (1924). On the dissociation of voluntary and emotional innervation in facial paralysis of central origin. *Brain, 47*, 22-35.

Moscovitch, M. (1983). The linguistic and emotional functions of the normal right hemisphere. In E. Perecman (Ed.), *Cognitive processing in the right hemisphere*. New York: Academic Press.

Natale, M., Gur, R. E., & Gur, R. C. (1983). Hemispheric asymmetries in processing emotional expressions. *Neuropsychologia, 21*, 555-565.

Oster, H., & Ekman, P. (1978). Facial behavior in child development. In W. A. Collins (Ed.), *Minnesota Symposium on child psychology* (Vol. 2, pp. 231-276). Hillsdale, NJ: Lawrence Erlbaum Associates.

Remillard, G. M., Anderman, F., Rhi-Sausi, A., & Robbins, N. M. (1977). Facial asymmetry in patients with temporal lobe epilepsy: A clinical sign useful in the lateralization of temporal epileptic foci. *Neurology, 27*, 109-114.

Reuter-Lorenz, P., & Davidson, R. J. (1981). Differential contributions of the two cerebral hemispheres to the perception of happy and sad faces. *Neuropsychologia, 19*, 609-613.

Reuter-Lorenz, P. A., Givis, R. P., & Moscovitch, M. (1983). Hemispheric specialization and the perception of emotion: Evidence from right-handers and from inverted and non-inverted left-handers. *Neuropsychologia, 21*, 687-692.

Rinn, W. E. (1984). The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin, 95*, 52-77.

Rossi, G. F., & Rosadini, G. (1967). Experimental analysis of cerebral dominance in man. In C. H. Millikan & F. L. Darley (Eds.), *Brain mechanisms underlying speech and language*. New York: Grune & Stratton.

Rothbart, M. K., & Posner, M. I. (1985). Temperament and the development of self-regulation. In L. C. Hartlage & F. C. Telzrow (Eds.), *The neuropsychology of individual differences: A developmental perspective* (pp. 93-123). New York: Plenum Press.

Rothbart, M. K., Taylor, S. B., & Tucker, D. M. (1989). Right-sided facial asymmetry in infant emotional expression. *Neuropsychologia, 27*, 675-687.

Sackeim, H. A., Greenberg, M. S., Weiman, A. L., Gur, R. C., Hungerbuhler, J. P., & Geschwind, N. (1982). Hemispheric asymmetry in the expression of positive and negative emotions. *Archives of Neurology, 39*, 210-218.

Sackeim, H. A., & Gur, R. C. (1978). Lateral asymmetry in intensity of emotional expression. *Neuropsychologia, 16*, 473-481.

Sackeim, H. A., & Gur, R. C. (1980). Asymmetry in facial expression. *Science, 209*, 834-836.

Safer, M. A. (1981). Sex and hemisphere differences in access to codes for processing emotional expressions and faces. *Journal of Experimental Psychology: General, 110*, 86-100.

Segalowitz, S. J. (1985). Hemispheric asymmetries in emotional faces. Paper presented at the [*International Neuropsychology Society* meeting, February 6-8, San Diego.

Silberman & Weingartner (1986). Hemispheric lateralization of functions related to emotion. *Brain and Cognition, 5*, 322-353.

Sroufe, L. A. (1979). Socioemotional development. In J. Osofsky (Ed.), *Handbook of infant development*. New York: John Wiley.

Strauss, E., & Moscovitch, M. (1981). Perceptual asymmetries in processing facial expression and facial identity. *Brain and Language, 13*, 308-322.

Terzian, H. (1964). Behavioral and EEG effects of intracarotid sodium amytal injection. *Acta Neurochirurgica, 12*, 230-239.

Tucker, D. M. (1981). Lateral brain function, emotion, and conceptualization. *Psychological Review, 89*, 19-46.

FOOTNOTES

†Also Wesleyan University, Middletown, CT.

††Wesleyan University.

[1]Performance level corrections such as the Phi coefficient (Kuhn, 1973) or λ (Bryden & Sprott, 1981) are neither necessary nor applicable with binary forced-choice data.

[2]However, this failure may be due, in part, to the manner in which those researchers paired their adult chimeras. Whereas we always paired the two normal orientation chimeras or the two mirror-reversed chimeras of a given infant for forced-choice judgments, Levy et al. (1983) had always paired a normal orientation chimera with its own mirror-image. Their approach may have masked their viewers' sensitivity to poser asymmetries. In an unpublished extension of their study we modified the pairings of the Levy et al. adult chimeras such that both members of each pair were in the same orientation (Best & Queen, 1987). Under those conditions we found a significant influence of poser asymmetries upon the viewers' perceptual field biases. Even in our extension of Levy et al., nonetheless, the influence of adult poser asymmetries upon the viewers' perceptual biases was very much smaller than that found with our infant faces. It should also be noted that the Levy et al. stimuli included only smiling expressions, which yielded the largest influence of poser asymmetries for infant expressions in the present study.

[3]The eye and mouth regions were not separated from one another until after the midline had been drawn on each face from the point midway between the internal canthi and the center of the philtrum, as in Experiment 1.

Appendix

SR #	Report Date	DTIC #	ERIC #
SR-21/22	January-June 1970	AD 719382	ED 044-679
SR-23	July-September 1970	AD 723586	ED 052-654
SR-24	October-December 1970	AD 727616	ED 052-653
SR-25/26	January-June 1971	AD 730013	ED 056-560
SR-27	July-September 1971	AD 749339	ED 071-533
SR-28	October-December 1971	AD 742140	ED 061-837
SR-29/30	January-June 1972	AD 750001	ED 071-484
SR-31/32	July-December 1972	AD 757954	ED 077-285
SR-33	January-March 1973	AD 762373	ED 081-263
SR-34	April-June 1973	AD 766178	ED 081-295
SR-35/36	July-December 1973	AD 774799	ED 094-444
SR-37/38	January-June 1974	AD 783548	ED 094-445
SR-39/40	July-December 1974	AD A007342	ED 102-633
SR-41	January-March 1975	AD A013325	ED 109-722
SR-42/43	April-September 1975	AD A018369	ED 117-770
SR-44	October-December 1975	AD A023059	ED 119-273
SR-45/46	January-June 1976	AD A026196	ED 123-678
SR-47	July-September 1976	AD A031789	ED 128-870
SR-48	October-December 1976	AD A036735	ED 135-028
SR-49	January-March 1977	AD A041460	ED 141-864
SR-50	April-June 1977	AD A044820	ED 144-138
SR-51/52	July-December 1977	AD A049215	ED 147-892
SR-53	January-March 1978	AD A055853	ED 155-760
SR-54	April-June 1978	AD A067070	ED 161-096
SR-55/56	July-December 1978	AD A065575	ED 166-757
SR-57	January-March 1979	AD A083179	ED 170-823
SR-58	April-June 1979	AD A077663	ED 178-967
SR-59/60	July-December 1979	AD A082034	ED 181-525
SR-61	January-March 1980	AD A085320	ED 185-636
SR-62	April-June 1980	AD A095062	ED 196-099
SR-63/64	July-December 1980	AD A095860	ED 197-416
SR-65	January-March 1981	AD A099958	ED 201-022
SR-66	April-June 1981	AD A105090	ED 206-038
SR-67/68	July-December 1981	AD A111385	ED 212-010
SR-69	January-March 1982	AD A120819	ED 214-226
SR-70	April-June 1982	AD A119426	ED 219-834
SR-71/72	July-December 1982	AD A124596	ED 225-212
SR-73	January-March 1983	AD A129713	ED 229-816
SR-74/75	April-September 1983	AD A136416	ED 236-753
SR-76	October-December 1983	AD A140176	ED 241-973
SR-77/78	January-June 1984	AD A145585	ED 247-626
SR-79/80	July-December 1984	AD A151035	ED 252-90
SR-81	January-March 1985	AD A156294	ED 257-159
SR-82/83	April-September 1985	AD A165084	ED 266-508
SR-84	October-December 1985	AD A168819	ED 270-831
SR-85	January-March 1986	AD A173677	ED 274-022
SR-86/87	April-September 1986	AD A176816	ED 278-066

SR-88	October-December 1986	PB 88-244256	ED 282-278
SR-89/90	January-June 1987	PB 88-244314	ED 285-228
SR-91	July-September 1987	AD A192081	**
SR-92	October-December 1987	PB 88-246798	**
SR-93/94	January-June 1988	PB 89-108765	**
SR-95/96	July-December 1988	PB 89-155329/AS	
SR-97/98	January-June 1989	PB 90-121161/AS	

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm Corporation (CMC)
3900 Wheeler Avenue
Alexandria, VA 22304-5110

In addition, Haskins Laboratories Status Report on Speech Research is abstracted in *Language and Language Behavior Abstracts*, P.O. Box 22206, San Diego, CA 92122

**Accession number not yet assigned

Haskins Laboratories Status Report on

Speech Research

SR-99/100
JULY-DECEMBER 1989

Contents

• Coarticulatory Organization for Lip-rounding in Turkish and English
Suzanne E. Boyce 1

• Long Range Coarticulatory Effects for Tongue Dorsum Contact in VCVCV Sequences
Daniel Recasens 19

• A Dynamical Approach to Gestural Patterning in Speech Production
Elliot L. Saltzman and Kevin G. Munhall 38

• Articulatory Gestures as Phonological Units
Catherine P. Browman and Louis Goldstein 69

• The Perception of Phonetic Gestures
Carol A. Fowler and Lawrence D. Rosenblum 102

• Competence and Performance in Child Language
Stephen Crain and Janet Dean Fodor 118

• Cues to the Perception of Taiwanese Tones
Hwei-Bing Lin and Bruno H. Repp 137

• Physical Interaction and Association by Contiguity in Memory for the Words and Melodies of Songs
Robert G. Crowder, Mary Louise Serafine, and Bruno H. Repp 148

• Orthography and Phonology: The Psychological Reality of Orthographic Depth
Ram Frost 162

• Phonology and Reading: Evidence from Profoundly Deaf Readers
Vicki L. Hanson 172

• Syntactic Competence and Reading Ability in Children
Shlomo Bentin, Avital Deutsch, and Isabelle Y. Liberman 180

• Effect of Emotional Valence in Infant Expressions upon Perceptual Asymmetries in Adult Viewers
Catherine T. Best and Heidi Fréya Queen 195

Appendix 211